(19) Europäisches Patentamt
European Patent Office
Office européen des brevets

(11) EP 0 692 134 B1

(12) **EUROPEAN PATENT SPECIFICATION**

(45) Date of publication and mention of the grant of the patent:
24.02.1999 Bulletin 1999/08

(21) Application number: **94911271.8**

(22) Date of filing: **31.03.1994**

(51) Int Cl.[6]: **G10L 5/06**

(86) International application number:
PCT/GB94/00704

(87) International publication number:
WO 94/23424 (13.10.1994 Gazette 1994/23)

(54) **SPEECH PROCESSING**

SPRACHVERARBEITUNG

TRAITEMENT DE LA PAROLE

(84) Designated Contracting States:
BE CH DE DK ES FR GB IT LI NL SE

(30) Priority: **31.03.1993 EP 93302538**
25.06.1993 EP 93304993

(43) Date of publication of application:
17.01.1996 Bulletin 1996/03

(73) Proprietor: **BRITISH TELECOMMUNICATIONS public limited company**
London EC1A 7AJ (GB)

(72) Inventor: **SMYTH, Samuel Gavin**
Felixstowe, Suffolk IP11 8UA (GB)

(74) Representative: **Roberts, Simon Christopher et al**
BT Group Legal Services,
Intellectual Property Department,
8th Floor, Holborn Centre
120 Holborn
London, EC1N 2TE (GB)

(56) References cited:
EP-A- 0 248 377
EP-A- 0 535 929
US-A- 4 980 918

- **ICASSP, vol.1, 23 May 1989, GLASGOW pages 667 - 670 S.C.AUSTIN ET AL. 'A unified direction mechanism for automatic speech recognition systems using hidden markov models'**

EP 0 692 134 B1

Description

[0001] The present invention relates to speech processing and in particular to a system for processing alternative parses of connected speech.

[0002] Speech processing includes speaker recognition, in which the identity of a speaker is detected or verified, and speech recognition, wherein a system may be used by anyone without requiring recogniser training, and so-called speaker dependent recognition, in which the users allowed to operate a system are restricted and a training phase is necessary to derive information from each allowed user. It is common in recognition processing to input speech data, typically in digital form, to a so-called front-end processor, which derives from the stream of input speech data a more compact, perceptually significant set of data referred to as a front-end feature set or vector. For example, speech is typically input via a microphone, sampled, digitised, segmented into frames of length 10-20ms (e.g. sampled at 8 kHz) and, for each frame, a set of coefficients is calculated. In speech recognition, the speaker is normally assumed to be speaking one of a known set of words or phrases. A stored representation of the word or phrase, known as a template or model, comprises a reference feature matrix of that word as previously derived from, in the case of speaker independent recognition, multiple speakers. The input feature vector is matched with the model and a measure of similarity between the two is produced.

[0003] Speech recognition (whether human or machine) is susceptible to error and may result in the misrecognition of words. If a word or phrase is incorrectly recognised, the speech recogniser may then offer another attempt at recognition, which may or may not be correct.

[0004] Various ways have been suggested for processing speech to select the best or alternative matches between input speech and stored speech templates or models. In isolated word recognition systems, the production of alternative matches is fairly straightforward: each word is a separate 'path' in a transition network representing the words to be recognised and the independent word paths join only at the final point in the network. Ordering all the paths exiting the network in terms of their similarity to the stored templates or the like will give the best and alternative matches.

[0005] In most connected recognition systems and some isolated word recognition systems based on connected recognition techniques however, it is not always possible to recombine all the paths at the final point of the network and thus neither the best nor alternative matches are directly obtainable from the information available at the exit point of the network. One solution to the problem of producing a best match is discussed in "Token Passing: a Simple Conceptual Model for Connected Speech Recognition Systems" by S. J. Young, N. H. Russell and J. H. S. Thornton 1989, which relates to passing packets of information, known as tokens, through a transition network. A token contains information relating to the partial path travelled as well as an accumulated score indicative of the degree of similarity between the input and the portion of the network processed thus far.

[0006] As described by Young et al, at each input of a frame of speech to a transition network, any tokens that are present at the input of a node are passed into the node and the current frame of speech matched within the word models associated with those nodes. New tokens then appear at the output of the nodes (having "travelled" through the model associated with the node). Only the best scoring token is then passed onto the inputs of the following nodes. When the end of speech has been signalled (by an external device such as a pause detector), a single token will be present at the final node. From this token the entire path through the network can be extracted by tracing back along the path by means of the previous path information contained within the token to provide the best match to the input speech.

[0007] The article "A unified direction mechanism for automatic speech recognition using Hidden Markov Models" by S.C. Austin and F. Fallside, ICASSP 1989, Vol. 1, pages 667-670, relates to a connected word speech recogniser which operates in a manner similar to that described by Young et al, as described above. A history relating to the progress of the recognition through the transition network is updated on exiting the word model. At the end of recognition, the result of recognition is derived from the history presented to the output which has the best score. Again only one history is possible for each path terminating at the final node.

[0008] Such known arrangements do not allow, for a given final node, for an alternative choice to be readily available at the output of the network.

[0009] In accordance with the invention a speech recognition apparatus comprises means for deriving a recognition feature vector from an input speech signal for each predetermined time frame; means for modelling expected input speech comprising a plurality of vocabulary nodes each of which has an associated word representation model and links between said vocabulary nodes; processing means for comparing the recognition feature vectors with the modelled input speech and for generating a path link for each node and time frame, said path links indicating the most likely prior sequence of vocabulary nodes for each vocabulary node and time frame, each path link comprising a field for storing an accumulated recognition score and a field for storing a reference to the most likely previous path link in the sequence; and means for indicating recognition of the input speech signal in dependence upon the comparison; characterised in that the processing means (351) is capable of processing more than one path link for at least one vocabulary node, other than the final node, in a single time frame.

[0010] Such an arrangement means that more than one incoming path link can be processed by a node in a single time frame and hence that more than one recognition result may be obtained.

[0011] The modelling means preferably comprises a transition network containing a plurality of noise nodes and vocabulary nodes which are associated with word represenation models. The processing means is capable of producing path links for each node comprising fields for storing a pointer to the previous path link, an accumulated score for a path, a pointer to a previous node and a time index for segmentation information. Preferably, the vocabulary nodes which may have more than one path link processed in a single time frame have more than one identical associated word representation model.

[0012] The provision that at least one of the vocabulary nodes other than the final node of the network has more than one associated word representation model allows the processor to process multiple paths for the same time frame and so allows more than one path link to be propagated across each inter-node link at each input frame. In effect, the invention creates multiple layers of a transition network along which several alternative paths may be propagated. The best scoring path may be used by the first model of a node, the next best by the second and so on until either parallel models or incoming paths run out.

[0013] In general terms "network" includes directed acyclic graphs (DAGs) and trees. A DAG is a network with no cycles and a tree is a network in which the only meeting of paths occurs conceptually right at the end of the network.

[0014] The term "word" here denotes a basic recognition unit, which may be a word but equally well may be a diphone, phoneme, allophone, etc. Recognition is the process of matching an unknown utterance with a predefined transition network, the network having been designed to be compatible with what a user is likely to say.

[0015] In order to identify the phrase that has been recognised, the apparatus may include means for tracing the path link back through the network.

[0016] Alternatively, the apparatus may also include means for assigning a signature to at least some of the nodes having associated word representation models and means for comparing the signature of each path, to determine the path with the best match to the input speech and that with the second best alternative match.

[0017] This arrangement allows for an alternative which is necessarily different in character to the best match and does not differ merely in segmentation or noise matches.

[0018] The word representation models may be Hidden Markov Models (HMMs) as described generally in British Telecom Technology Journal, April 1988, Vol, 6, no. 2, page 105: Cox, "Hidden Markov Models for automatic speech recognition: theory and application", templates, Dynamic Time Warping models or any other suitable word representation model. The processing which occurs within a model is irrelevant as far as this invention is concerned.

[0019] It is not necessary for all the nodes having associated word models to have a signature assigned to them. Depending on the structure of the transition network, it may be sufficient only to assign signatures to those nodes which appear before a decision point within a network. A decision point as used herein relates to a point in the network which has more than one incoming path.

[0020] Partial paths may be examined at certain decision points in the network, certain constraints being imposed at these decision points so that only paths conforming to the constraints are propagated, as described in the applicants' International patent application filed on 31st March 1994 entitled "Connected Speech Recognition", No WO/23425, published 13th October 1994.. Each decision point is associated with a set of valid signatures and any path links with signatures that are not in the set are discarded.

[0021] The accumulated signature may be used to identify the complete path, resulting in extra efficiency of operation as the path links need not be traversed to determine the path identity, and the partial path information of the token may not be generated at all. In this case the signature field must be large enough to identify all paths uniquely.

[0022] For efficient operation of the apparatus according to the invention, the signal processing of path signatures is preferably carried out in a single operation to increase processing speed.

[0023] Other aspects and preferred embodiments of the invention are as disclosed and claimed herein, with advantages that will be apparent hereafter.

[0024] The invention will now be described further, by way of example only, with reference to the accompanying drawings in which:

Figure 1 shows schematically the employment of a recognition processor according to the invention in a telecommunications environment;

Figure 2 is a block diagram showing schematically the functional elements of a recognition processor according to the invention;

Figure 3 is a block diagram indicating schematically the components of a classifier forming part of Figure 2;

Figure 4 is block diagram showing schematically the structure of a sequence parser forming part of the embodiment of Figure 2;

Figure 5 shows schematically the content of a field within a store forming part of Figure 4;

Figure 6 is a schematic representation of one embodiment of a transition network applicable with the processor

of the sequence parser of Figure 4;
Figure 7a shows a node of a network and Figure 7b shows a path link as employed according to the invention;
Figures 8 to 10 show the progression of path links through the network of Figure 6;
Figure 11 is a schematic representation of a second embodiment of a transition network of a apparatus according to the invention;
Figure 12 is a schematic representation of a third embodiment of a transition network of an apparatus according to the invention.

[0025] Referring to Figure 1, a telecommunications system including speech recognition generally comprises a microphone 1, typically forming part of a telephone handset, a telecommunications network (typically a public switched telecommunications network (PSTN)) 2, a recognition processor 3, connected to receive a voice signal from the network 2, and a utilising apparatus 4 connected to the recognition processor 3 and arranged to receive therefrom a voice recognition signal, indicating recognition or otherwise of a particular word or phrase, and to take action in response thereto. For example, the utilising apparatus 4 may be a remotely operated banking terminal for effecting banking transactions.

[0026] In many cases, the utilising apparatus 4 will generate an auditory response to the speaker, transmitted via the network 2 to a loudspeaker 5 typically forming a part of the subscriber handset.

[0027] In operation, a speaker speaks into the microphone 1 and an analog speech signal is transmitted from the microphone 1 into the network 2 to the recognition processor 3, where the speech signal is analysed and a signal indicating identification or otherwise of a particular word or phrase is generated and transmitted to the utilising apparatus 4, which then takes appropriate action in the event of recognition of the speech.

[0028] Typically, the recognition processor needs to acquire data concerning the speech against which to verify the speech signal, and this data acquisition may be performed by the recognition processor in a second mode of operation in which the recognition processor 3 is not connected to the utilising apparatus 4, but receives a speech signal from the microphone 1 to form the recognition data for that word or phrase. However, other methods of acquiring the speech recognition data are also possible.

[0029] Typically, the recognition processor 3 is ignorant of the route taken by the signal from the microphone 1 to and through the network 2; any one of a large variety of types and qualities of receiver handset. Likewise, within the network 2, any one of a large variety of transmission paths may be taken, including radio links, analog and digital paths and so on. Accordingly, the speech signal Y reaching the recognition processor 3 corresponds to the speech signal S received at the microphone 1, convolved with the transfer characteristics of the microphone 1, link to network 2, channel through the network 2, and link to the recognition processor 3, which may be lumped and designated by a single transfer characteristic H.

[0030] Referring to Figure 2, the recognition processor 3 comprises an input 31 for receiving speech in digital form (either from a digital network or from an analog to digital converter), a frame processor 32 for partioning the succession of digital samples into a succession of frames of contiguous samples; a feature extractor 33 for generating from a frame of samples a corresponding feature vector; a classifier 34 receiving the succession of feature vectors and operating on each with a plurality of model states, to generate recognition results; a sequencer 35 which is arranged to receive the classification results form the classifier 34 and to determine the predetermined utterance to which the sequence of classifier output indicates the greatest similarity; and an output port 38 at which a recognition signal is supplied indicating the speech utterance which has been recognised.

Frame Generator 32

[0031] The frame generator 32 is arranged to receive speech samples at a rate of, for example, 8,000 samples per second, and to form frames comprising 256 contiguous samples, at a frame rate of 1 frame every 16ms. Preferably, each frame is windowed (i.e. the samples towards the edge of the frame are multiplied by predetermined weighting constants) using, for example, a Hamming window to reduce spurious artifacts, generated by the frames edges. In a preferred embodiment, the frames are overlapping (for example by 50%) so as to ameliorate the effects of the windowing.

Feature Extractor 33

[0032] The feature extractor 33 receives frames from the frame generator 32 and generates, in each case, a set or vector of features. The features may, for example, comprise cepstral coefficients (for example, LPC cepstral coefficients or mel frequency cepstral coefficients as described in "On the Evaluation of Speech Recognisers and Databases using a Reference System", Chollet & Gagnoulet, 1982 proc. IEEE p2026), or differential values of such coefficients comprising, for each coefficient, the differences between the coefficient and the corresponding coefficient value in the

preceding vector, as described in "On the use of Instantaneous and Transitional Spectral Information in Speaker Recognition", Soong & Rosenberg, 1988 IEEE Trans. on Acoustics, Speech and Signal Processing Vol 36 No. 6 p871. Equally, a mixture of several types of feature coefficient may be used.

[0033] The feature extractor 33 outputs a frame number, incremented for each successive frame. The output of the feature extractor 33 is also passed to an end pointer 36, the output of which is connected to the classifier 34. The end pointer 36 detects the end of speech and various types are well known in this field.

[0034] The frame generator 32 and feature extractor 33 are, in this embodiment, provided by a single suitably programmed digital signal processor (DSP) device (such as the Motorola DSP 56000, or the Texas Instruments TMS C 320) or similar device.

Classifier 34

[0035] Referring to Figure 3, in this embodiment, the classifier 34 comprises a classifying processor 341 and a state memory 342.

[0036] The state memory 342 comprises a state field 3421, 3422, ..., for each of the plurality of speech states. For example, each allophone to be recognised by the recognition processor comprises 3 states, and accordingly 3 state fields are provided in the state memory 342 for each allophone.

[0037] The classification processor 34 is arranged to read each state field within the memory 342 in turn, and calculate for each, using the current input feature coefficient set, the probability that the input feature set or vector corresponds to the corresponding state.

[0038] Accordingly, the output of the classification processor is a plurality of state probabilities P, one for each state in the state memory 342, indicating the likelihood that the input feature vector corresponds to each state.

[0039] The classifying processor 341 may be a suitably programmed digital signal processing (DSP) device, may in particular be the same digital signal processing device as the feature extractor 33.

Sequencer 35

[0040] Referring to Figure 4, the sequencer 35 in this embodiment comprises a state sequence memory 352, a parsing processor 351, and a sequencer output buffer 354.

[0041] Also provided is a state probability memory 353 which stores, for each frame processed, the state probabilities output by the classifier processor 341. The state sequence memory 352 comprises a plurality of state sequence fields 3521, 3522, ..., each corresponding to a word or phase sequence to be recognised consisting of a string of allophones.

[0042] Each state sequence in the state sequence memory 352 comprises, as illustrated in Figure 5, a number of states P_1, P_2, ... P_N (where N is a multiple of 3) and, for each state, two probabilities; a repeat probability (P_{i1}) and a transition probability to the following state (P_{i2}). The states of the sequence are a plurality of groups of three states each relating to a single allophone. The observed sequence of states associated with a series of frames may therefore comprise several repetitions of each state P_i in each state sequence model 3521 etc; for example:

Frame Number	1	2	3	4	5	6	7	8	9	. . .	Z	Z+1
State	P1	P1	P1	P2	P2	P2	P2	P2	P2	. . .	Pn	Pn

[0043] The parsing processor 351 is arranged to read, at each frame, the state probabilities output by the classifier processor 341, and the previous stored state probabilities in the state probability memory 353, and to calculate the most likely path of states to date over time, and to compare this with each of the state sequences stored in the state sequence memory 352.

[0044] The calculation employs the well known HMM, as discussed in the above referenced Cox paper. Conveniently, the HMM processing performed by the parsing processor 351 uses the well . known Viterbi algorithm. The parsing processor 351 may, for example, be a microprocessor such as the Intel(TM) i-486(TM) microprocessor or the Motorola(TM) 68000 microprocessor, or may alternatively be a DSP device (for example, the same DSP device as is employed for any of the preceding processors).

[0045] Accordingly for each state sequence (corresponding to a word, phrase or other speech sequence to be recognised) a probability score is output by the parser processor 351 at each frame of input speech. For example the state sequences may comprise the names in a telephone directory. When the end of the utterance is detected, a label signal a indicating the most probable state sequence is output from the parsing processor 351 to the output port 38, to indicate that the corresponding name, word or phrase has been recognised.

[0046] The parsing processor 351 comprises a network which is specifically configured to recognise certain phrases or words for example a string of digits.

[0047] Figure 6 shows a simple network for recognising a string of words, in this case either a string of four words or a string of three. Each node 12 of the network is associated with a word representation model 13, for example a HMM, which is stored in a model list. Several nodes can be associated with each model and each node includes a pointer to its associated model (as can be seen in Figures 6 and 7a). In order to produce a best match and a single alternative parse, the final node 14 is associated with two models so allowing this node to process two paths. If n parses are required, the final node 14 of the network is associated with n identical word models.

[0048] As shown in Figure 7b, a path link 15 contains information relating to a pointer to the previous path link, an accumulated score, a pointer to the node previously exited and a time index. At the start of an utterance, an empty path link 15' is inserted into the first node 16 as shown in Figure 8. The first node now contains a path link and is therefore active whereas the remaining nodes are inactive. At each clock tick (i.e. with each incoming frame of speech) any active nodes accumulate a score in their path link.

[0049] If the first model can match, say, a minimum of seven frames of speech, then at the seventh clock pulse a path link 15" is output from the first node with the score for matching the seven frames to the model and pointers to the entry path link and the node just matched. The path link is fed to all of the following nodes 12, as shown in Figure 9. Now the first three nodes are active. The input frame of speech is then matched in the models associated with the active nodes and new path links outputted.

[0050] This processing continues, with the first node 16 producing further path links as its model matches increasingly longer parts of the utterance and the succeeding nodes performing similar calculations.

[0051] When the input speech has been processed as far as the final node 18 of the network, path links from each 'branch' of the network may be presented to this node 18. If, at any given time frame, there is a single path link (i.e. only one of the parallel paths has been completed) that path link is taken to be the best (and only) match and is processed by the final node 18. However, if there are two path links presented to the final node 18, both are processed by that node, since the final node 18 is able to process more than one path. The output path links are continuously updated at each frame of speech. When the utterance is completed there will be two path links 15"' at the output of the network, as shown in Figure 10 (from which the pointers to previous path links and nodes have been excluded for the sake of clarity).

[0052] The full path can be found by following the pointers to the previous path links and the nodes on the recognised path (and hence the input speech deemed to be recognised) can be identified by looking at the pointers to the nodes exited.

[0053] Figure 11 represents a second embodiment of a network configured to recognise strings of three digits. The grey nodes 22 are null nodes in the network; the white nodes are active nodes which may be divided into vocabulary nodes 24 with associated word representation models (not shown), for matching incoming speech and noise nodes 25 which represent arbitrary noise.

[0054] If all of the active nodes 24, 25 after and including the third null node 22' are each capable of having three paths for each time frame (i.e. each vocabulary node 24 is associated with three word representation models), the output of the network will comprise path links relating to the three top scoring paths of the system. As described with reference to Figures 8 to 10, the three paths can be found by following, for each path, the pointers to the previous path links. The nodes on the paths (and hence the input speech deemed to be recognised) can be identified by looking at the pointers to the exited nodes.

[0055] In a further development of the invention, the path links may be augmented with signatures which represent the significant nodes of the network. These significant nodes may, for example, include all vocabulary nodes 24. In the embodiment of Figure 11, each vocabulary node 24 is assigned a signature, for example the nodes representing the digit 1 are assigned a signature '1', the nodes 24" representing the digit 2 are assigned a signature '2' and so on.

[0056] At the start of parsing, a single empty path link is passed into a network entry node 26. Since this is a null node, the path link is passed to the next node, a noise node 25. The input frame is matched in the noise model (not shown) of this node and an updated path link produced at the output. This path link is then passed to the next active nodes i.e.the first vocabulary nodes 24 having an associated model (not shown). Each vocabulary node 24 processes the frame of speech in its associated word model and produces an updated path link. The signature field of the path link is also updated. At the end of each time frame, the updated path links are sorted to retain the three (n) top scoring paths which have different signature fields. A list ordered by score is maintained with the added constraint that accumulated signatures are unique: if a second path link with the same signature enters, the better of the two is retained. The list contains only the top "n" different paths, the rest being ignored.

[0057] The n path links are propagated through the next null node 22' to the following noise node 25 and vocabulary nodes 24", each of which are associated with three identical word representation models. After this, model processing takes place, resulting in the updating of the lists of path links and the extending of the paths into further nodes 24"', 25. It should be clear that the signature fields of the path links are not updated after processing by the null nodes 22 or the noise nodes 25 since these nodes do not have assigned signatures.

[0058] The path links are propagated along paths which pass through the remaining active nodes to produce, at an

output node 28, up to three paths links indicating the relative scores and signatures, for example 1 2 1, of the paths taken through the network. The path links are continuously updated until the end of speech is detected (for example by an external device such as a pause detector or, until a time out is reached). At this point, the pointers or the accumulated signatures of the path links at the output node 28 are examined to determine the recognition results.

[0059] For example, presuming that the following three path links are presented to the output node 28 at some time instant:

	SCORE	SIGNATURE
A	10	1 2 2
B	9	1 2 2
C	7	1 3 2

Path A, the highest scoring path, is the best match. However, although path B has the second best score, it would be rejected as an alternative parse since its signature, and hence the speech deemed to be recognised, is the same as path A. Path C would therefore be retained as the second best parse.

[0060] If the strings to be recognised have more structure than that discussed above, for example spelt names, signatures need only be assigned to nodes just before decision points, rather than at every vocabulary node. Figure 12 shows a network for recognising the spelling of the names "Phil", "Paul" and "Peter". For simplicity, no noise is illustrated. The square nodes 44 indicate where the signature should be augmented.

[0061] The system can distinguish between the 'PHI' and 'PAU' paths at the 'L' node because the signatures of the path links created at the previous nodes are different. The following node 47 will be able to distinguish between all three independent paths since the signatures of the square nodes 44 are different. Only the 'L' node and the final noise node 48 need to be associated with more than one identical word model, so that these models are capable of having more than one path for a single time frame.

[0062] In all cases, each network illustrating the speech to be recognised requires analysis to determine which nodes are to be assigned signatures. In addition the network is configured to be compatible with what a user is likely to say.

[0063] Savings in memory size and processing speed may be achieved by limiting the signatures that a node will propagate. For instance, say the only valid input speech to a recogniser having the network of Figure 6 is the four following numbers only: 111, 112, 121, 211. Certain nodes within the network are associated with a set of valid signatures and a path will only be propagated by such a 'constrained' node if a path link having one of these signatures is presented. To achieve this, the signature fields of the path links entering a constrained node, eg third null node 22', are examined. If the signature field contains a signature other than 1 or 2, the path link is discarded and the path is not propagated any further. If an allowable path link is presented, it is passed on to the next node. The next constrained node is the null node 22" after the next vocabulary nodes. This null node is constrained to only propagate path links having a signature 11, 12 or 21. The null node 22"' after the next vocabulary nodes is constrained to only propagate path links having the signature 111, 112, 121 or 211. Such an arrangement significantly reduces the necessary processing and allows for a saving in the memory capacity of the apparatus. Only some of the nodes at decision points in the network need to be so constrained. In practice, a 32 bit signature has proved to be suitable for sequences of up to 9 digits. A 64 bit signature appears suitable for a 12 character alphanumeric string.

[0064] End of speech detection and various other aspects of speech recognition relevant to the present invention are more fully set out in the applicants' International Patent Application filed on 25th March 1994 entitled "Speech Recognition", No WO 94/22131, published 29th September 1994.

[0065] In the above described embodiments, recognition processing apparatus suitable to be coupled to a telecommunications exchange has been described. However, in another embodiment, the invention may be embodied on simple apparatus connected to a conventional subscriber station (mobile or fixed) connected to the telephone network; in this case, analog to digital conversion means may be provided for digitising the incoming analog telephone signal.

Claims

1. A speech recognition apparatus comprising:

   means for deriving a recognition feature vector from an input speech signal for each predetermined time frame;
   means for modelling expected input speech comprising a plurality of vocabulary nodes each of which vocabulary nodes has an associated word representation model and the model having links between said vocabulary nodes;
   processing means for comparing the recognition feature vectors with the modelled input speech and for gen-

erating a path link for each node and time frame, said path links indicating the most likely prior sequence of vocabulary nodes for each vocabulary node and time frame, each path link comprising a field for storing an accumulated recognition score and a field for storing a reference to the most likely previous path link in the sequence; and
means for indicating recognition of the input speech signal in dependence upon the comparison;

characterised in that the processing means (351) is capable of processing more than one path link for at least one vocabulary node, other than the final node, in a single time frame.

2. A speech recognition apparatus according to Claim 1 characterised in that the at least one of the vocabulary nodes are associated with more than one identical word representation model.

3. A speech recognition apparatus according to Claim 2 characterised in that the word representation models are Hidden Markov Models.

4. A speech recognition apparatus according to any one of Claims 1,2 or 3 characterised in that all the vocabulary nodes have signatures assigned to them.

5. A speech recognition apparatus according to any one of claims 1,2 or 3 characterised in that only the vocabulary nodes occurring before a decision point have signatures assigned to them.

6. A speech recognition apparatus according to either Claim 4 or 5 characterised in that the path links include an accumulated signature.

7. A speech recognition apparatus according to any one of Claims 4, 5 or 6 characterised in that at least some of the nodes are constrained only to propagate path links having certain predetermined signatures.

8. A speech recognition apparatus according to any one of Claims 4 to 7 characterised in that the recognition indicating means includes means for comparing the score and signature of the path links to determine the path with the best match to the input connected speech and those with the next best alternative matches.

9. A method of speech recognition comprising:

deriving a recognition feature vector from an input speech signal for each predetermined time frame;
modelling expected input speech;
comparing the feature data with the modelled input speech by generating a network containing a plurality of vocabulary nodes associated with word representation models and generating a path link for each node and time frame, the path link indicating the most likely prior sequence of vocabulary nodes for each vocabulary node and time frame, each path link comprising a field for storing an accumulated recognition score and a field for storing a reference to the most likely previous path link in the sequence;
indicating recognition of the speech in dependence upon the comparison characterised in that more than one path link is processed in a single time frame for at least one vocabulary node other than the final node.

10. A method according to Claim 9 characterised in that the at least one of the vocabulary nodes is associated with more than one identical word representation model.

11. A method according to Claim 10 characterised in that the at least one of the vocabulary nodes is associated with a number of identical word representation models equal to the number of desired recognition results.

12. A method according to either Claim 10 or 11 characterised in that the scores of the path links are compared at each decision point of the network, only the n top scoring path links being propagated to the next nodes(s).

13. A method according to any one of claims 10, 11 or 12 characterised by assigning signatures to all the vocabulary nodes.

14. A method according to any one of claims 10, 11 or 12 characterised by only assigning signatures to vocabulary nodes occurring before a decision point in the network.

15. A method according to either Claim 13 or 14 when appended to Claim 12 characterised in that the signatures of the path links are also compared, only path links having different signatures being propagated to the next node(s).

16. A method according to any one of Claims 13, 14 or 15 characterised by constraining at least some nodes only to pass path links having certain predetermined signatures in their signature fields.

17. A method according to any one of Claims 9 to 16 characterised in that the input speech signal deemed to be recognised is determined by tracing the path links back through the network.

18. A method according to any one of Claims 13 to 16 characterised in that the input speech signal deemed to be recognised is determined by the accumulated signature of each path link.

19. A method according to any one of Claims 10 to 18 characterised in that the best scoring path link is processed by the first word representation model of a vocabulary node, the next best by the second and so on until either parallel models or incoming path links run out.

Patentansprüche

1. Spracherkennungssystem, mit :

einer Einrichtung zum Ableiten eines Erkennungsmerkmalsvektors aus einem eingegebenen Sprachsignal für jeden vorgegebenen Zeitrahmen;
einer Einrichtung zum Modellieren einer erwarteten eingegebenen Sprache, die mehrere Vokabularknoten enthält, wovon jeder ein zugeordnetes Wortdarstellungsmodell besitzt, das seinerseits Verbindungen zwischen den Vokabularknoten aufweist;
einer Verarbeitungseinrichtung zum Vergleichen der Erkennungsmerkmalsvektoren mit der modellierten eingegebenen Sprache und zum Erzeugen einer Wegverbindung für jeden Knoten und jeden Zeitrahmen, wobei die Wegverbindungen die wahrscheinlichste vorhergehende Sequenz von Vokabularknoten für jeden Vokabularknoten und jeden Zeitrahmen angeben, wobei jede Wegverbindung ein Feld zum Speichern einer akkumulierten Erkennungstrefferliste und ein Feld zum Speichern einer Bezugnahme auf die wahrscheinlichste vorhergehende Wegverbindung in der Sequenz enthält; und
einer Einrichtung, die die Erkennung des eingegebenen Sprachsignals in Abhängigkeit vom Vergleich angibt;

dadurch gekennzeichnet, daß die Verarbeitungseinrichtung (351) in einem einzigen Zeitrahmen mehr als eine Wegverbindung für wenigstens einen vom Endknoten verschiedenen Vokabularknoten verarbeiten kann.

2. Spracherkennungsvorrichtung nach Anspruch 1, dadurch gekennzeichnet, daß der wenigstens eine der Vokabularknoten mehr als einem identischen Wortdarstellungsmodell zugeordnet ist.

3. Spracherkennungsvorrichtung nach Anspruch 2, dadurch gekennzeichnet, daß die Wortdarstellungsmodelle Hidden-Markow-Modelle sind.

4. Spracherkennungsvorrichtung nach irgendeinem der Ansprüche 1, 2 oder 3, dadurch gekennzeichnet, daß sämtliche Vokabularknoten ihnen zugewiesene Signaturen besitzen.

5. Spracherkennungsvorrichtung nach irgendeinem der Ansprüche 1, 2 oder 3, dadurch gekennzeichnet, daß nur diejenigen Vokabularknoten, die vor einem Entscheidungspunkt auftreten, ihnen zugewiesene Signaturen besitzen.

6. Spracherkennungsvorrichtung nach Anspruch 4 oder Anspruch 5, dadurch gekennzeichnet, daß die Wegverbindungen eine akkumulierte Signatur enthalten.

7. Spracherkennungsvorrichtung nach irgendeinem der Ansprüche 4, 5, oder 6, dadurch gekennzeichnet, daß wenigstens einige der Knoten in der Weise beschränkt sind, daß von ihnen nur Wegverbindungen mit bestimmten vorgegebenen Signaturen ausgehen.

8. Spracherkennungsvorrichtung nach irgendeinem der Ansprüche 4 bis 7, dadurch gekennzeichnet, daß die Erken-

nungsangabeeinrichtung eine Einrichtung zum Vergleichen der Trefferliste und der Signatur der Wegverbindungen enthält, um den Weg mit der besten Übereinstimmung mit der eingangsseitigen Sprache und jene Wege mit den nächstbesten alternativen Übereinstimmungen zu bestimmen.

9. Verfahren zur Spracherkennung, enthaltend:

Ableiten eines Erkennungsmerkmalsvektors aus einem eingegebenen Sprachsignal für jeden vorgegebenen Zeitrahmen;
Modellieren einer erwarteten Eingangssprache;
Vergleichen der Merkmalsdaten mit der modellierten Eingangssprache durch Erzeugen eines Netzes, das mehrere Vokabularknoten enthält, denen Wortdarstellungsmodelle zugeordnet sind, und durch Erzeugen einer Wegverbindung für jeden Knoten und jeden Zeitrahmen, wobei die Wegverbindung die wahrscheinlichste vorhergehende Sequenz aus Vokabularknoten für jeden Vokabularknoten und jeden Zeitrahmen angibt, wobei jede Wegverbindung ein Feld zum Speichern einer akkumulierten Erkennungstrefferliste und ein Feld zum Speichern einer Bezugnahme auf die wahrscheinlichste vorhergehende Wegverbindung in der Sequenz enthält;
Angeben einer Erkennung der Sprache in Abhängigkeit vom Vergleich,

dadurch gekennzeichnet, daß in einem einzigen Zeitrahmen mehr als eine Wegverbindung für wenigstens einen von dem Endknoten verschiedenen Vokabularknoten verarbeitet werden.

10. Verfahren nach Anspruch 9, dadurch gekennzeichnet, daß dem wenigstens einen Vokabularknoten mehr als ein identisches Wortdarstellungsmodell zugeordnet ist.

11. Verfahren nach Anspruch 10, dadurch gekennzeichnet, daß dem wenigstens einem Vokabularknoten eine Anzahl identischer Wortdarstellungsmodelle zugeordnet ist, die gleich der Anzahl der gewünschten Erkennungsergebnisse ist.

12. Verfahren nach irgendeinem der Ansprüche 10 oder 11, dadurch gekennzeichnet, daß die Trefferlisten der Wegverbindungen mit jedem Entscheidungspunkt des Netzes verglichen werden, wobei nur die Wegverbindungen mit bester Trefferliste zu dem/den nächsten Knoten fortgeführt werden.

13. Verfahren nach irgendeinem der Ansprüche 10, 11 oder 12, gekennzeichnet durch die Zuweisung von Signaturen an sämtliche Vokabularknoten.

14. Verfahren nach irgendeinem der Ansprüche 10, 11 oder 12, dadurch gekennzeichnet, daß lediglich denjenigen Vokabularknoten, die vor einem Entscheidungspunkt im Netz auftreten, Signaturen zugewiesen werden.

15. Verfahren nach irgendeinem der Ansprüche 13 oder 14 in Verbindung mit Anspruch 12, dadurch gekennzeichnet, daß die Signaturen der Wegverbindungen ebenfalls verglichen werden, wobei nur Wegverbindungen mit unterschiedlichen Signaturen zu dem/den nächsten Knoten fortgeführt werden.

16. Verfahren nach irgendeinem der Ansprüche 13, 14 oder 15, gekennzeichnet durch die Beschränkung wenigstens einiger Knoten in der Weise, daß sie nur Wegverbindungen mit bestimmten vorgegebenen Signaturen in ihren Signaturfeldern weiterleiten.

17. Verfahren nach irgendeinem der Ansprüche 9 bis 16, dadurch gekennzeichnet, daß das eingegebene Sprachsignal, das erkannt werden soll, durch Zurückverfolgen der Wegverbindungen durch das Netz bestimmt wird.

18. Verfahren nach irgendeinem der Ansprüche 13 bis 16, dadurch gekennzeichnet, daß das eingegebene Sprachsignal, das erkannt werden soll, durch die akkumulierte Signatur jeder Wegverbindung bestimmt wird.

19. Verfahren nach irgendeinem der Ansprüche 10 bis 18, dadurch gekennzeichnet, daß die Wegverbindung mit bester Trefferliste durch das erste Wortdarstellungsmodell eines Vokabularknotens verarbeitet wird, die nächstbeste durch das zweite u.s.w., bis entweder parallele Modelle oder ankommende Wegverbindung ausgehen.

Revendications

1. Dispositif de reconnaissance de la parole comprenant :

un moyen destiné à obtenir un vecteur de caractéristiques de reconnaissance à partir d'un signal vocal en entrée pour chaque trame temporelle prédéterminée,
un moyen destiné à modéliser la parole en entrée prévue, comprenant un certain nombre de noeuds de vocabulaire, chacun des noeuds de vocabulaire comprenant un modèle de représentation de mots associé et le modèle comportant des liaisons entre lesdits noeuds de vocabulaire,
un moyen de traitement destiné à comparer les vecteurs de caractéristiques de reconnaissance à la parole en entrée modélisée et à générer une liaison de trajet pour chaque noeud et une trame temporelle, lesdites liaisons de trajets indiquant la séquence antérieure la plus probable des noeuds de vocabulaire pour chaque noeud de vocabulaire et chaque trame temporelle, chaque liaison de trajet comprenant un champ destiné à mémoriser une évaluation accumulée de la reconnaissance et un champ destiné à mémoriser une référence sur la liaison de trajet précédente la plus probable dans la séquence, et
un moyen destiné à indiquer la reconnaissance du signal vocal en entrée suivant la comparaison,

caractérisé en ce que le moyen de traitement (351) est capable de traiter plus d'une liaison de trajet pour au moins un noeud de vocabulaire, autre que le noeud final, dans une seule trame temporelle.

2. Dispositif de reconnaissance de la parole selon la revendication 1, caractérisé en ce que le au moins un noeud des noeuds de vocabulaire est associé à plus d'un modèle identique de représentation de mots.

3. Dispositif de reconnaissance de la parole selon la revendication 2, caractérisé en ce que les modèles de représentation de mots sont des Modèles de Markov Cachés.

4. Dispositif de reconnaissance de la parole selon l'une quelconque des revendications 1, 2 ou 3, caractérisé en ce que tous les noeuds de vocabulaire se voient affecter des signatures.

5. Dispositif de reconnaissance de la parole selon l'une quelconque des revendications 1, 2 ou 3, caractérisé en ce que seuls les noeuds de vocabulaire apparaissant avant un point de décision se voient affecter des signatures.

6. Dispositif de reconnaissance de la parole selon soit la revendication 4, soit la revendication 5, caractérisé en ce que les liaisons de trajets comprennent une signature accumulée.

7. Dispositif de reconnaissance de la parole selon l'une quelconque des revendications 4, 5 ou 6, caractérisé en ce qu'au moins certains des noeuds sont contraints à ne propager que des liaisons de trajets présentant certaines signatures prédéterminées.

8. Dispositif de reconnaissance de la parole selon l'une quelconque des revendications 4 à 7, caractérisé en ce que le moyen d'indication de reconnaissance comprend un moyen destiné à comparer l'évaluation et la signature des liaisons de trajets afin de déterminer le trajet présentant la meilleure mise en correspondance avec la parole reliée en entrée et ceux présentant les mises en correspondance auxiliaires les meilleures suivantes.

9. Procédé de reconnaissance de la parole comprenant :

l'obtention d'un vecteur de caractéristiques de reconnaissance à partir d'un signal vocal en entrée pour chaque trame temporelle prédéterminée,
la modélisation de la parole en entrée prévue,
la comparaison des données de caractéristiques avec la parole en entrée modélisée en générant un réseau contenant un certain nombre de noeuds de vocabulaire associés à des modèles de représentation de mots, et la génération d'une liaison de trajet pour chaque noeud et trame temporelle, la liaison de trajet indiquant la séquence antérieure la plus probable de noeuds de vocabulaire pour chaque noeud de vocabulaire et chaque trame temporelle, chaque liaison de trajet comprenant un champ destiné à mémoriser une évaluation de reconnaissance accumulée et un champ destiné à mémoriser une référence vers la liaison de trajet précédente la plus probable dans la séquence,
l'indication de la reconnaissance de la parole suivant la comparaison, caractérisé en ce que plus d'une liaison de trajet est traitée dans une seule trame temporelle pour au moins un noeud de vocabulaire autre que le

noeud final.

10. Procédé selon la revendication 9, caractérisé en ce que le au moins un noeud parmi les noeuds de vocabulaire est associé à plus d'un modèle identique de représentation de mots.

11. Procédé selon la revendication 10, caractérisé en ce que le au moins un noeud parmi les noeuds de vocabulaire est associé à un nombre de modèles de représentation de mots identiques égal au nombre des résultats de reconnaissance désirés.

12. Procédé selon soit la revendication 10, soit la revendication 11, caractérisé en ce que les évaluations des liaisons de trajets sont comparées au niveau de chaque point de décision du réseau, seules les n liaisons de trajets présentant la meilleure évaluation étant propagées vers le ou les noeuds suivants.

13. Procédé selon l'une quelconque des revendications 10, 11 ou 12, caractérisé par l'affectation de signatures à tous les noeuds de vocabulaire.

14. Procédé selon l'une quelconque des revendications 10, 11 ou 12, caractérisé par le fait de n'affecter des signatures qu'à des noeuds de vocabulaire apparaissant avant un point de décision dans le réseau.

15. Procédé selon soit la revendication 13, soit la revendication 14, lorsqu'elle dépend de la revendication 12, caractérisé en ce que les signatures des liaisons de trajets sont également comparées, seules les liaisons de trajets présentant des signatures différentes étant propagées vers le ou les noeuds suivants.

16. Procédé selon l'une quelconque des revendications 13, 14 ou 15, caractérisé par le fait de contraindre au moins certains des noeuds à ne transmettre que des liaisons de trajets présentant certaines signatures prédéterminées dans leur champ de signature.

17. Procédé selon l'une quelconque des revendications 9 à 16, caractérisé en ce que le signal vocal en entrée estimé être reconnu est déterminé en remontant les liaisons de trajets au travers du réseau.

18. Procédé selon l'une quelconque des revendications 13 à 16, caractérisé en ce que le signal vocal en entrée estimé être reconnu est déterminé par la signature accumulée de chaque liaison de trajet.

19. Procédé selon l'une quelconque des revendications 10 à 18, caractérisé en ce que la liaison de trajet présentant la meilleure évaluation est traitée par le premier modèle de représentation de mots d'un noeud de vocabulaire, la seconde meilleure par le second et ainsi de suite jusqu'à l'épuisement, soit des modèles parallèles, soit des liaisons de trajet entrantes.

Fig.1

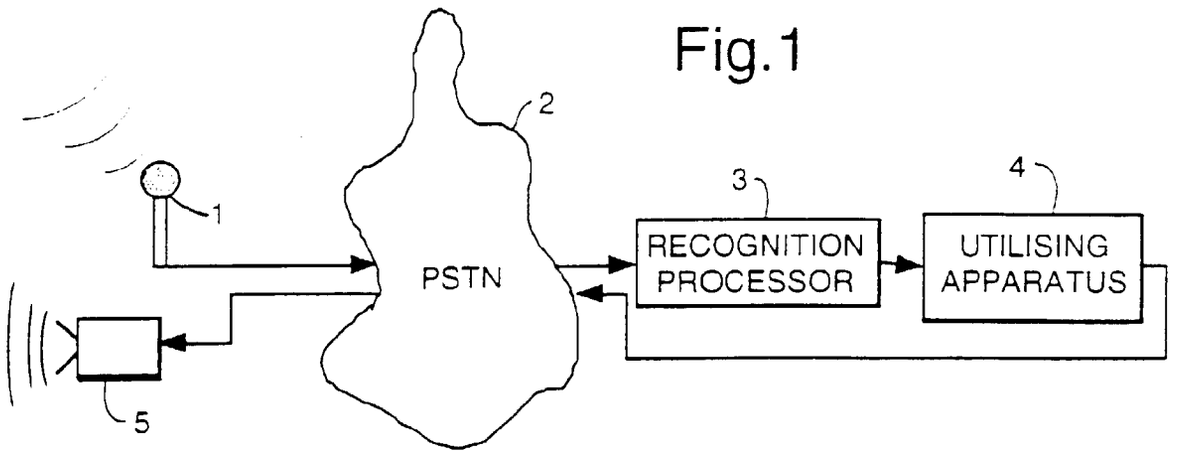

1
2
3
4
5
PSTN
RECOGNITION PROCESSOR
UTILISING APPARATUS

Fig.2

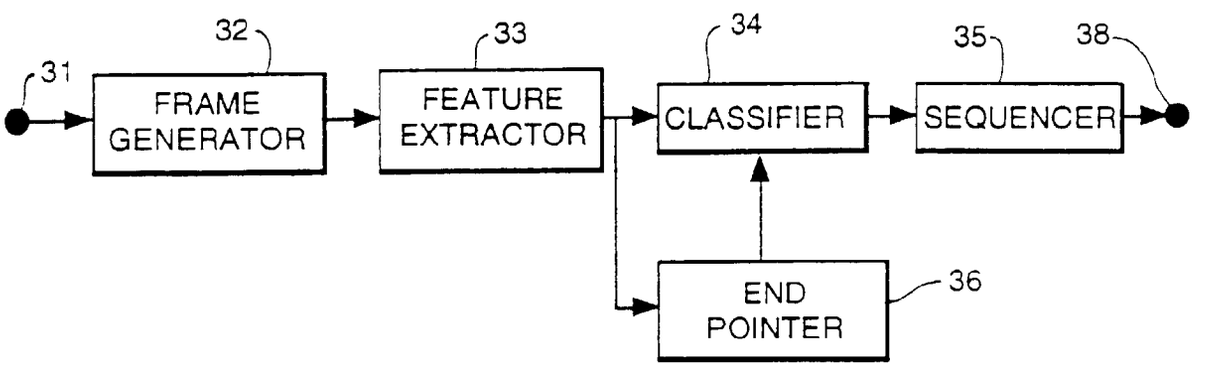

31
32
33
34
35
36
38
FRAME GENERATOR
FEATURE EXTRACTOR
CLASSIFIER
SEQUENCER
END POINTER

Fig.3

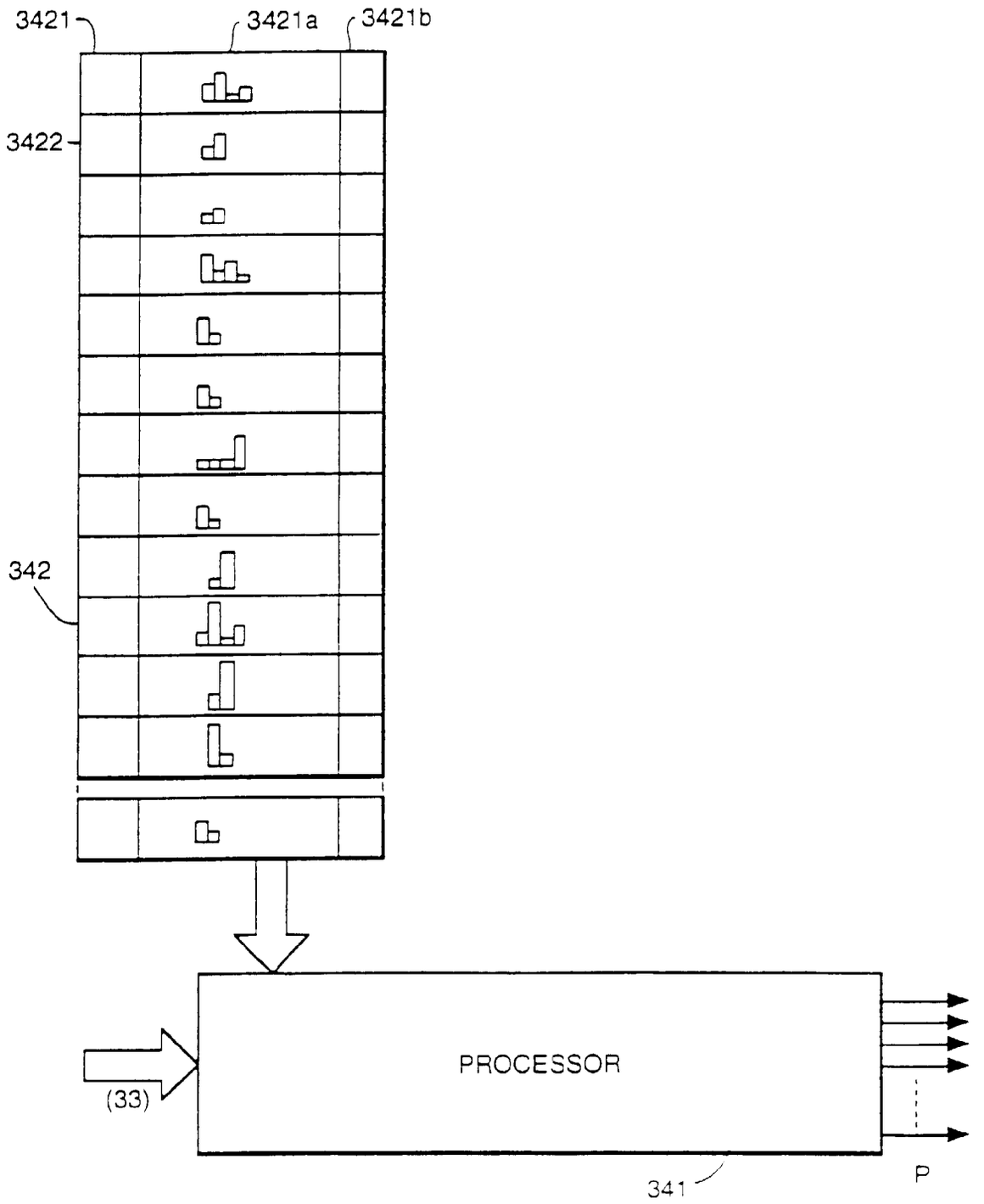


Fig.4.

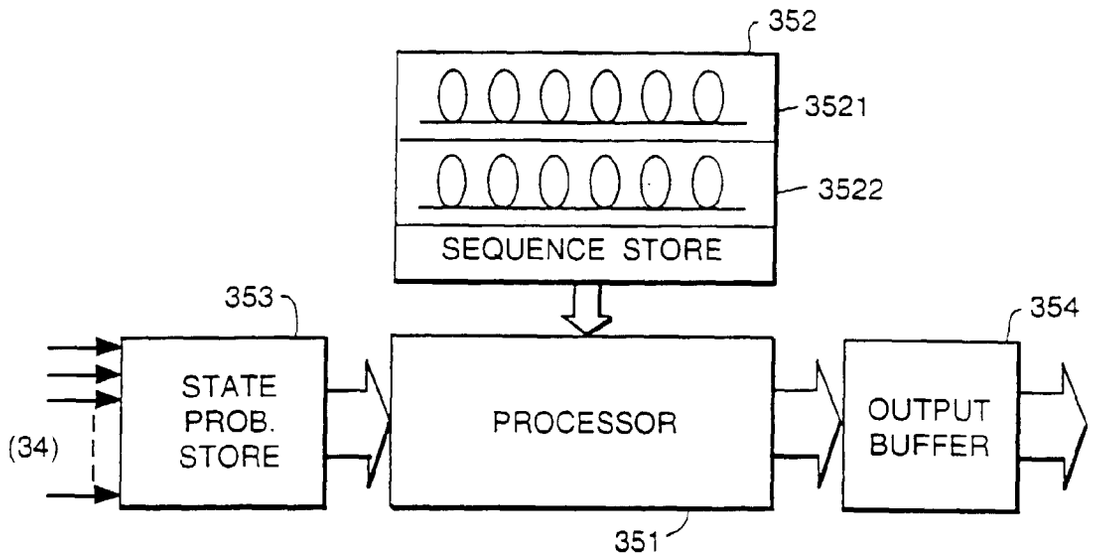

352
3521
3522
SEQUENCE STORE
353
354
STATE
PROB.
STORE
(34)
PROCESSOR
OUTPUT
BUFFER
351

Fig.5.

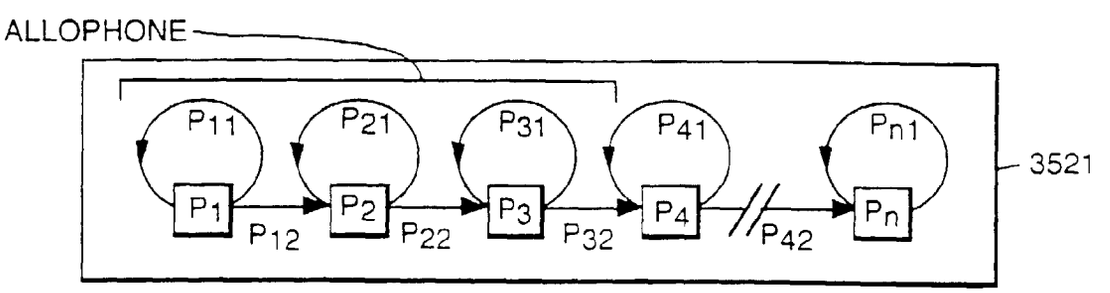

ALLOPHONE
P11
P21
P31
P41
Pn1
P1
P2
P3
P4
Pn
P12
P22
P32
P42
3521

Fig.6.

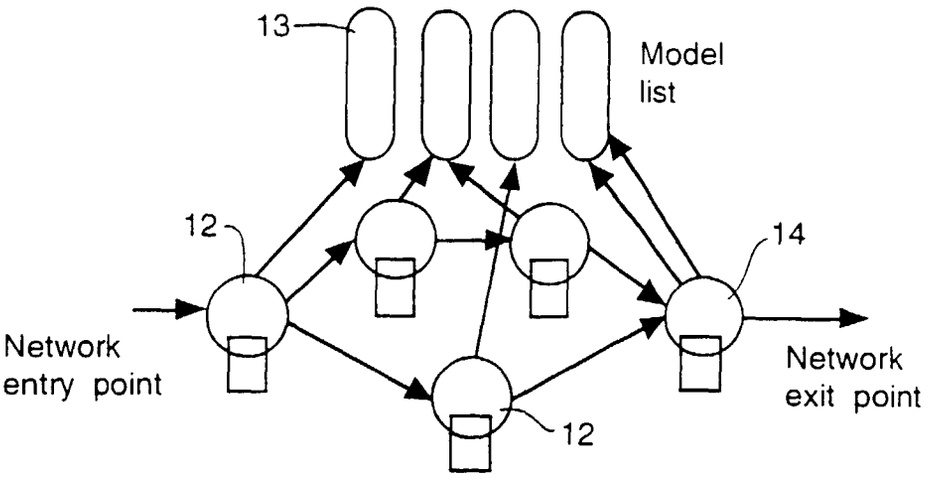


Fig.7a

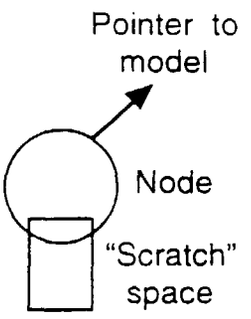


Fig.7b

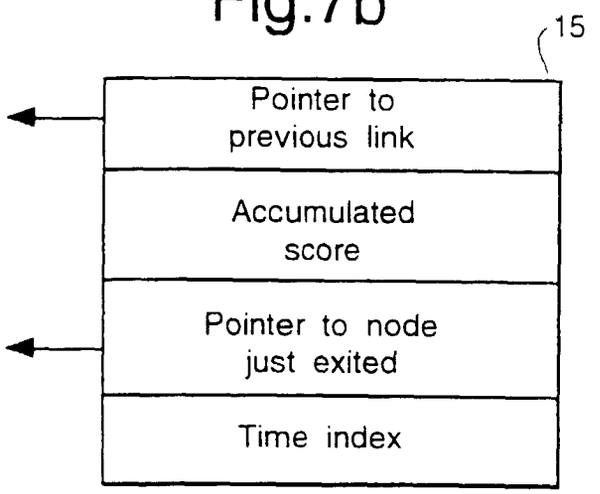


Fig.8.

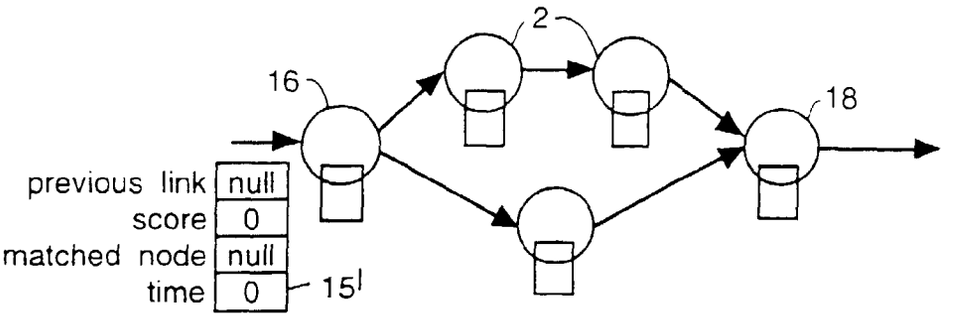


Fig.9.

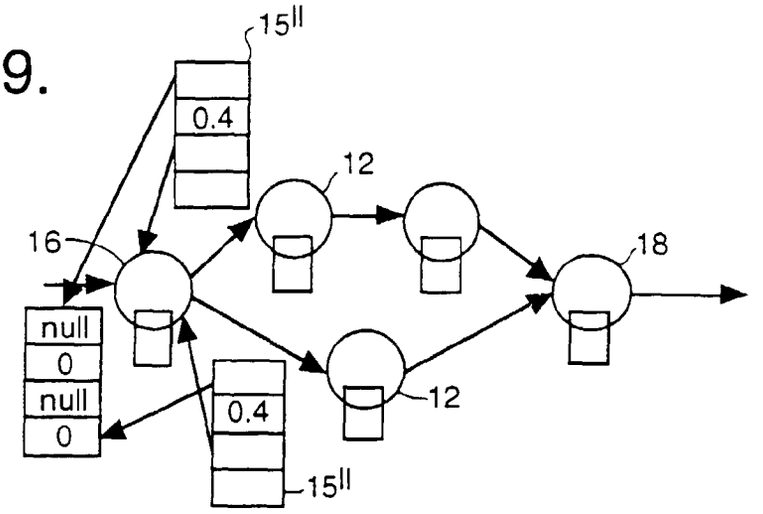


Fig.10.

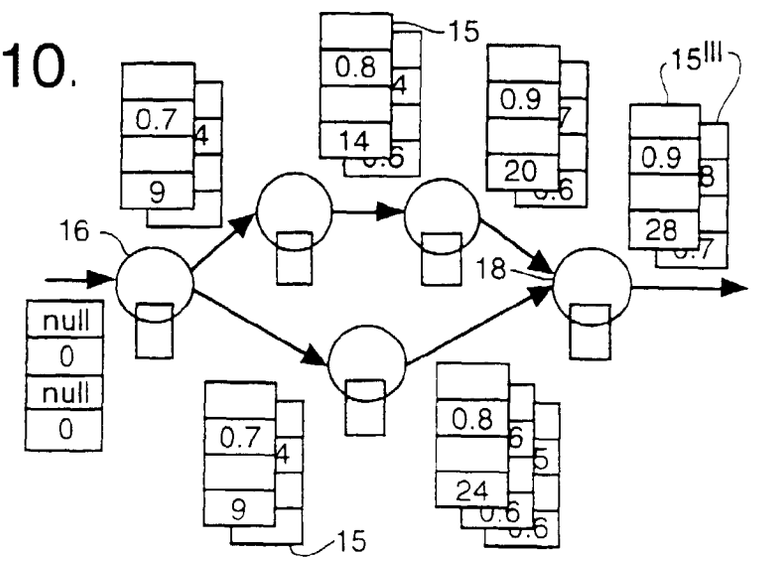


Fig.11.

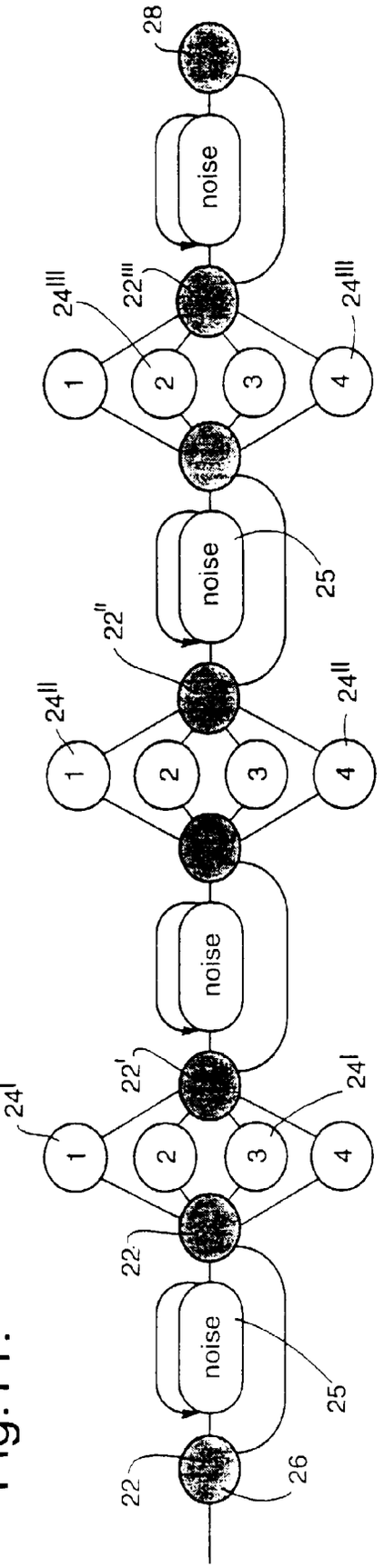
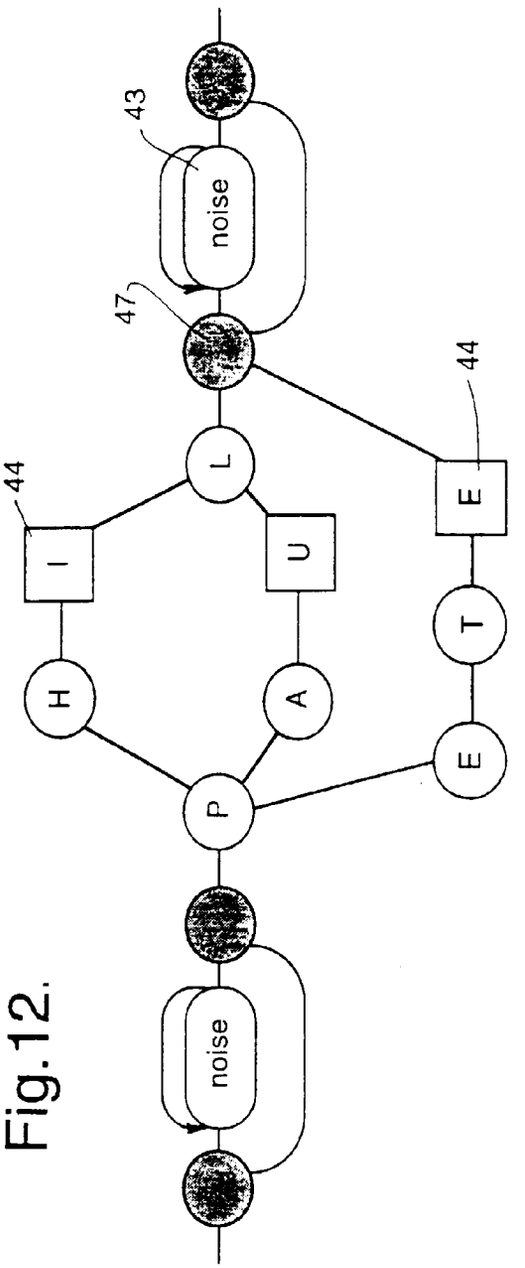

noise
noise
noise
noise
1
2
3
4
22
22'
22''
22'''
24'
24''
24'''
24''''
25
26
28

Fig.12.


noise
H
I
L
U
A
P
E
T
E
44
43
47
44