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

(11) EP 1 220 202 A1

(12) EUROPEAN PATENT APPLICATION

(43) Date of publication:
03.07.2002 Bulletin 2002/27

(51) Int Cl.[7]: **G10L 19/00**

(21) Application number: **00440335.8**

(22) Date of filing: **29.12.2000**

(84) Designated Contracting States:
AT BE CH CY DE DK ES FI FR GB GR IE IT LI LU MC NL PT SE TR
Designated Extension States:
AL LT LV MK RO SI

(71) Applicant: **ALCATEL**
75008 Paris (FR)

(72) Inventors:
- **Sienel, Jürgen**
  D-71229 Leonberg (DE)
- **Kopp, Dieter**
  D-75428 Illingen (DE)
- **Reher, Norbert**
  D-70806 Kornwestheim (DE)

(74) Representative: **van Bommel, Jan Peter et al**
Alcatel
Intellectual Property Department, Stuttgart
70430 Stuttgart (DE)

(54) System and method for coding and decoding speaker-independent and speaker-dependent speech information

(57) Coders ans decoders which do not distinguish between speaker-independent signals and speaker-dependent signals in voice/speech signals communicate inefficiently. By separately coding speaker-independent signals (for example phonemes) and speaker-dependent signals (for example prosody, amplitude) in the coder into two streams of coded signals, and transmitting both streams of coded signals possibly in a multiplex way to the decoder, which separately decodes both streams of coded signals into speaker-independent signals and speaker-dependent signals, the entire system is more efficient. Preferably, in the decoder, after having decoded one of both streams of coded signals, the result is used for decoding the other stream. Further, said speaker-dependent signals may be divided into time-independent signals and time-dependent signals, for further optimisation. Finally, said system can be a distributed speech recognition system, with the coder performing the preprocessing (prerecognition) and a network between coder and decoder performing the final processing (final recognition).

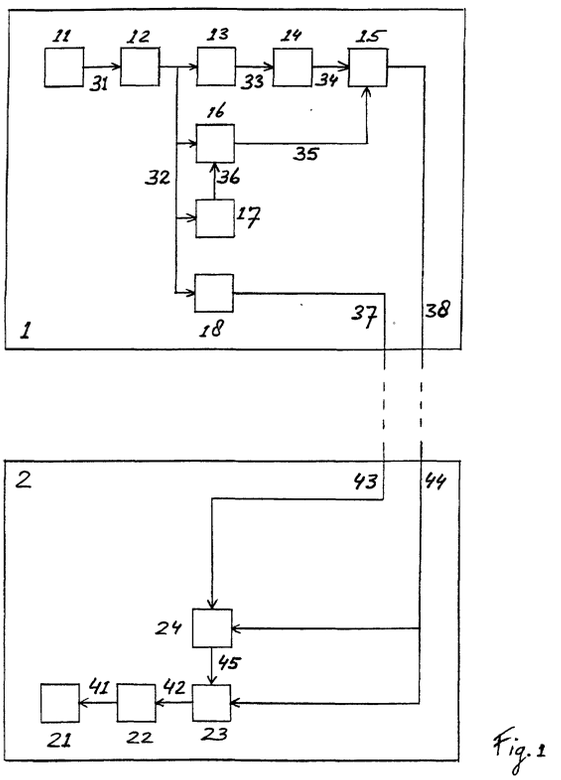

Fig. 1

EP 1 220 202 A1

Description

[0001] The invention relates to a system comprising a coder for coding a voice/speech signal into at least one coded signal and comprising a decoder for decoding at least one further signal.

[0002] Such a system is known from EP 0 718 819, in which a first coder is used for coding vocally produced audio, like spoken words, singing and other vocal utterances, and in which a second coder is used for coding non-vocally produced audio, like music. Said known system further comprises a first decoder and a second decoder for decoding purposes. The at least one further signal may either correspond entirely, or partly, or not at all with the at least one coded signal.

[0003] Such a system is disadvantageous, inter alia, due to voice/speech like said spoken words being coded and decoded in an inefficient way.

[0004] It is an object of the invention, inter alia, to provide a system described in the preamble, which is more efficient.

[0005] Thereto, the system according to the invention is characterised in that said system comprises a system processor system for processing a speaker-independent signal of said voice/speech signal and in response generating a first coded signal and for processing a speaker-dependent signal of said voice/speech signal and in response generating a second coded signal and for processing a first further signal and in response generating a speaker-independent signal and for processing a second further signal and in response generating a speaker-dependent signal.

[0006] Said speaker-independent signal of said voice/speech signal corresponds for example with phoneme information like letters and parts of words which are coded into said first coded signal, and said a speaker-dependent signal of said voice/speech signal corresponds for example with prosody information like a user-volume and user-voice-frequencies which are coded into said second coded signal. Said first further signal for example corresponds (either entirely, or partly, or not at all) with said first coded signal and then is related to said phoneme information like letters and parts of words which are decoded into said speaker-independent signal, and said second further signal for example corresponds (either entirely, or partly, or not at all) with said prosody information like a user-volume and user-voice-frequencies which are decoded into said speaker-dependent signal.

[0007] By introducing the separately processing of speaker-independent signals and speaker-dependent signals, the efficiency of the system is increased.

[0008] The invention is based on the insight, inter alia, that voice/speech coding/decoding processes comprise redundancies which can be removed without reducing the quality.

[0009] The invention solves the problem, inter alia, of providing a more efficient system, by introducing the separately processing.

[0010] Where EP 0 718 819 distinguishes between vocally produced audio, like spoken words, singing and other vocal utterances, and non-vocally produced audio, like music, the system according to the invention distinguishes between speaker-independent signals withing voice/speech signals and speaker-dependent signals withing said voice/speech signals.

[0011] The invention further relates to a coder for coding a voice/speech signal into at least one coded signal.

[0012] The coder according to the invention is characterised in that said coder comprises a processor system for processing a speaker-independent signal of said voice/speech signal and in response generating a first coded signal and for processing a speaker-dependent signal of said voice/speech signal and in response generating a second coded signal.

[0013] A first embodiment of the coder according to the invention is characterised in that said speaker-dependent signal comprises a time-independent part and a time-dependent part, with said processor system processing said time-independent part and in response generating a third coded signal and processing said time-dependent part and in response generating a fourth coded signal.

[0014] By introducing, with respect to the speaker-dependent signals, the separately processing of time-independent and time-dependent signals, the efficiency is further increased. Said third coded signal and said fourth coded signal together correspond with (at least a part of) said second coded signal.

[0015] A second embodiment of the coder according to the invention is characterised in that said coder forms part of a distributed speech recognition system, with said processor system preprocessing said speaker-dependent signal.

[0016] By locating this coder in a distributed speech recognition system, said system becomes more efficient.

[0017] The invention yet further relates to a decoder for decoding at least one coded signal.

[0018] The decoder according to the invention is characterised in that said decoder comprises a processor system for processing a first coded signal and in response generating a speaker-independent signal and for processing a second coded signal and in response generating a speaker-dependent signal.

[0019] A first embodiment of the decoder according to the invention is characterised in that at least one of both speaker-independent signal and speaker-dependent signal is generated in dependence of the other one.

[0020] By using a result (the speaker-independent signals or the speaker-dependent signals) of one of the decoding processes and/or by using an input signal (one of both coded signals) of one of the decoding processes for the other decoding process, the decoder has an increased efficiency.

[0021] A second embodiment of the decoder accord-

ing to the invention is characterised in that said decoder decodes a third coded signal and a fourth coded signal, with said processor system processing said third coded signal and in response generating a time-independent part of said speaker-dependent signal and processing said fourth coded signal and in response generating a time-dependent part of said speaker-dependent signal.

[0022] A third embodiment of the decoder according to the invention is characterised in that said decoder forms part of a distributed speech recognition system, with said processor system final processing said second coded signal.

[0023] The invention also relates to a coding method for coding a voice/speech signal into at least one coded signal.

[0024] The coding method according to the invention is characterised in that said method comprises the steps of processing a speaker-independent signal of said voice/speech signal and in response generating a first coded signal and of processing a speaker-dependent signal of said voice/speech signal and in response generating a second coded signal.

[0025] The invention yet also relates to a decoding method for decoding at least one coded signal.

[0026] The decoding method according to the invention is characterised in that said method comprises the steps of processing a first coded signal and in response generating a speaker-independent signal and of processing a second coded signal and in response generating a speaker-dependent signal.

[0027] Embodiments of both methods according to the invention correspond with embodiments of the coder and/or decoder according to the invention.

[0028] Embodiments of the system according to the invention correspond with embodiments of the coder and/or decoder according to the invention.

[0029] EP 0 718 819, in which a first coder is used for coding vocally produced audio, like spoken words, singing and other vocal utterances, and in which a second coder is used for coding non-vocally produced audio, like music, does not disclose the separately processing, within its first coder for coding vocally produced audio, of speaker-independent signals and speaker-dependent signals both forming part of voice/speech signals, and does not disclose that said speaker-independent signals and speaker-dependent signals, within its decoder for decoding the coded vocally produced audio, are separately processed and separately decoded, possibly with at least one of the decoding processes being dependent upon the other one. US 5,012,518 discloses a low bit-rate speech coder, and US 5,388,181 discloses a digital audio compression system. Neither one of these documents discloses the system according to the invention, the coder according to the invention, the decoder according to the invention or the methods according to the invention.

[0030] All references, including references cited with respect to these references, are considered to be incorporated.

[0031] The invention will be further explained at the hand of an embodiment described with respect to a drawing, whereby

figure 1 discloses a system according to the invention comprising a coder according to the invention and a decoder according to the invention.

[0032] Coder 1 according to the invention as shown in figure 1 comprises a voice/speech receiver 11 like a microphone for receiving voice/speech signals of which an output is coupled via a coupling 31 to an input of sampler 12, of which an output is coupled via a coupling 32 to an input of first analyser 13 for determining for example speaker-dependent frequency information of said voice/speech signals and to an input of second analyser 16 for determining for example speaker-dependent prosody information in said voice/speech signals and to an input of third analyser 17 for determining for example speaker-dependent volume information of said voice/speech signals and to a phoneme unit 18 for determining for example speaker-independent phoneme information in said voice/speech signals. An output of first analyser 13 is coupled via a coupling 33 to an input of a fourth analyser 14 for determining for example vocal tract information, of which an output is coupled via a coupling 34 to an input of a combining unit 15. An output of third analyser 17 is coupled via a coupling 36 to a further input of second analyser 16, of which an output is coupled via a coupling 35 to a further input of combining unit 15. An output of combining unit 15 is coupled via a coupling 38 to decoder 2, and an output of phoneme unit 18 is coupled via a coupling 37 to decoder 2.

[0033] Decoder 2 according to the invention as shown in figure 1 comprises a first converting unit 23 for converting for example a combination of frequency information, prosody information and volume information into a first vocal signal, of which an input is coupled to a coupling 44 which is coupled to coupling 38, and of which a further input is coupled via a coupling 45 to an output of second converting unit 24 for converting for example speaker-independent phoneme information into a second vocal signal, of which an input is coupled to a coupling 43 which is coupled to coupling 37, and of which a further input is coupled to a coupling 44. An output of first converting unit 23 is coupled via a coupling 42 to an input of a third converting unit 22 for performing for example a digital/analog conversion, of which an output is coupled to an input of a voice/speech generator 21, like for example a loudspeaker.

[0034] The coder 1 according to the invention and the decoder 2 according to the invention as shown in figure 1 function as follows.

[0035] According to a first embodiment, coder 1 forms part of a first terminal, and decoder 2 forms part of a second terminal. Voice/speech originating from a first user at said first terminal is received by receiver 11, and then sampled (and possibly coded) by sampler 12. The result is supplied to first analyser 13 for determining (for

example calculating) for example (speaker-dependent, time-independent) frequency information (like one or more basic frequencies) of said voice/speech signals and to second analyser 16 for determining (for example calculating) for example (speaker-dependent, time-dependent) prosody information (like intonations) in said voice/speech signals and to third analyser 17 for determining (calculating) for example (speaker-dependent, time-dependent) volume information (like amplitudes) of said voice/speech signals and to a phoneme unit for determining (calculating) for example (speaker-independent) phoneme information (like letters, parts of words) in said voice/speech signals. Fourth analyser 14 for determining (calculating) for example vocal tract information receives said frequency information directly (unamendedly) or indirectly (after being processed), and combining unit 15 combines the results coming from fourth analyser 14 and second analyser 16, which has used the result coming from third analyser 17. The output signal generated by combining unit 15 is supplied directly (unamendedly) or indirectly (after being processed) via (wired or wireless) couplings 38 and 44 (and for example one or more networks, switches, routers, bridges, mobile switching centers, basestations, etc.) to decoder 2 and corresponds with a speaker-dependent part of said voice/speech signals in coded form. The output signal generated by phoneme unit 18 is supplied directly (unamendedly) or indirectly (after being processed) via (wired or wireless) couplings 37 and 43 (and for example one or more networks, switches, routers, bridges, mobile switching centers, basestations, etc.) to decoder 2 and corresponds with a speaker-independent part of said voice/speech signals in coded form.

[0036] Second converting unit 24 in decoder 2 receives both output signals directly (unamendedly) or indirectly (after being processed), and generates in response the second vocal signal which is supplied to first converting unit 23, which receives said second vocal signal as well as said output signal arrived via coupling 44, each one of both directly (unamendedly) or indirectly (after being processed), and generates the first vocal signal, which via said third converting unit 22 and said voice/speech generator 21 is converted into voice/speech signals destined for a second user at said second terminal.

[0037] As a result of using said coder 1 and decoder 2, the highest coding/decoding quality has been combined with the highest transmission efficiency (lowest bit rates - lowest capacity needed).

[0038] According to a first alternative to said first embodiment, in coder 1 couplings 37 and 38 are combined, for example multiplexed by a multiplexer not shown, in which case decoder 2 will comprise a demultiplexer for demultiplexing etc.

[0039] According to a second alternative to said first embodiment, coder 1 and decoder 2 each comprise a processor and a memory both not shown and coupled to at least one of said receiver 11, sampler 12, analysers 13,14,16,17, phoneme unit 18, combining unit 15, converting units 22,23,24 and generator 21. As a result, in said coder and decoder, intelligence is centralised.

[0040] Therefore, couplings 37-43 and 38-44 may be realised via different wires or one wire, different fibers or one fiber, different wireless links or one wireless link, and via different channels or one channel, different timeslots or same timeslots, different codes or same codes etc.

[0041] According to a second embodiment, (at least a part of) said coder 1 is located in a terminal, and (at least a part of) said decoder 2 is located in a network, or vice versa. This will for example be necessary in case old fashioned terminals not comprising these coders/decoders and novel hightech terminals comprising these coders/decoders are both used, whereby in the network for each possibly communication via an old fashioned terminal a coder/decoder needs to be available.

[0042] According to a third embodiment, said coder 1 forms part of a distributed speech recognition (DSR) system, whereby in said coder 1 said speaker-dependent signal is preprocessed, and final processing is done generally in said network or exceptionally in said decoder 2. This for example corresponds with at least a part of at least one function performed by at least one of said four analysers 13,14,16,17 is shifted generally into the network or exceptionally into the decoder 2, and/or at least a part of at least one function performed by at least one of said two converting units 23,24 is shifted into the network.

[0043] According to a fourth embodiment, in coder 1 at least some of blocks 12,13,14,15, 16,17,18 represent functions performed by a processor system running one or more software programs thereby using a memory, and in decoder 2 at least some of blocks 22,23,24 represent functions performed by a processor system running one or more software programs thereby using a memory.

[0044] According to a fifth embodiment, in decoder 2 there is an option of amending and/or inserting speaker-dependent information, by for example providing converting unit 23 with a yet further input for receiving additional speaker-dependent information and/or for receiving amending information for amending said speaker-dependent information and/or by for example interrupting coupling 44 via for example an interruptor not shown (like a switch having a conductive and a non-conductive state) located either in coder 1 and/or in decoder 2. As a result, a user may select the voice speaking to said user.

[0045] According to a first alternative to said fifth embodiment, coder 1 and/or decoder 2 are provided with one or more memories not shown for storing for example the signals present at one or more of the couplings 37, 38, 43 and 44, to allow generation of these signals later in time, under control of a user, a terminal and/or a network(-unit).

[0046] According to a second alternative to said fifth

embodiment, for example the signals present at one or more of the couplings 37, 38, 43 and 44, are stored in one or more memories not shown and located outside coder 1 and/or decoder 2, to allow generation of these signals later in time, under control of a user, a terminal and/or a network(-unit).

[0047] According to a third alternative to said fifth embodiment, coder 1 and/or decoder 2 and/or a terminal and/or a network-unit are provided with for example a phoneme recogniser not shown and/or a memory not shown for verification purposes to allow verification of phoneme signals sent and/or received earlier, under control of a user, a terminal and/or a network(-unit), for example for checking Wallstreet orders and/or (trans) actions, by generating (for example unamendable) phoneme signals stored before. Therefore, a method of doing business (comprising a step of generating phoneme signals stored before, possibly via said recogniser) is not to be excluded.

[0048] All embodiments are just embodiments and do not exclude other embodiments not shown and/or described. All alternatives are just alternatives and do not exclude other alternatives not shown and/or described. Each (part of an) embodiment and/or each (part of an) alternative can be combined with any other (part of an) embodiment and/or any other (part of an) alternative. Terms like "in response to K" and "in dependence of L" and "for doing M" do not exclude that there could be a further "in response to N" and a further "in dependence of O" and a further "for doing P" etc.

Claims

1. System comprising a coder for coding a voice/speech signal into at least one coded signal and comprising a decoder for decoding at least one further signal, **characterised in that** said system comprises a system processor system for processing a speaker-independent signal of said voice/speech signal and in response generating a first coded signal and for processing a speaker-dependent signal of said voice/speech signal and in response generating a second coded signal and for processing a first further signal and in response generating a speaker-independent signal and for processing a second further signal and in response generating a speaker-dependent signal.

2. Coder for coding a voice/speech signal into at least one coded signal, **characterised in that** said coder comprises a processor system for processing a speaker-independent signal of said voice/speech signal and in response generating a first coded signal and for processing a speaker-dependent signal of said voice/speech signal and in response generating a second coded signal.

3. Coder according to claim 2, **characterised in that** said speaker-dependent signal comprises a time-independent part and a time-dependent part, with said processor system processing said time-independent part and in response generating a third coded signal and processing said time-dependent part and in response generating a fourth coded signal.

4. Coder according to claim 2 or 3, **characterised in that** said coder forms part of a distributed speech recognition system, with said processor system preprocessing said speaker-dependent signal.

5. Decoder for decoding at least one coded signal, **characterised in that** said decoder comprises a processor system for processing a first coded signal and in response generating a speaker-independent signal and for processing a second coded signal and in response generating a speaker-dependent signal.

6. Decoder according to claim 5, **characterised in that** at least one of both speaker-independent signal and speaker-dependent signal is generated in dependence of the other one.

7. Decoder according to claim 5 or 6, **characterised in that** said decoder decodes a third coded signal and a fourth coded signal, with said processor system processing said third coded signal and in response generating a time-independent part of said speaker-dependent signal and processing said fourth coded signal and in response generating a time-dependent part of said speaker-dependent signal.

8. Decoder according to claim 5, 6 or 7, **characterised in that** said decoder forms part of a distributed speech recognition system, with said processor system final processing said second coded signal.

9. Coding method for coding a voice/speech signal into at least one coded signal, **characterised in that** said method comprises the steps of processing a speaker-independent signal of said voice/speech signal and in response generating a first coded signal and of processing a speaker-dependent signal of said voice/speech signal and in response generating a second coded signal.

10. Decoding method for decoding at least one coded signal, **characterised in that** said method comprises the steps of processing a first coded signal and in response generating a speaker-independent signal and of processing a second coded signal and in response generating a speaker-dependent signal.

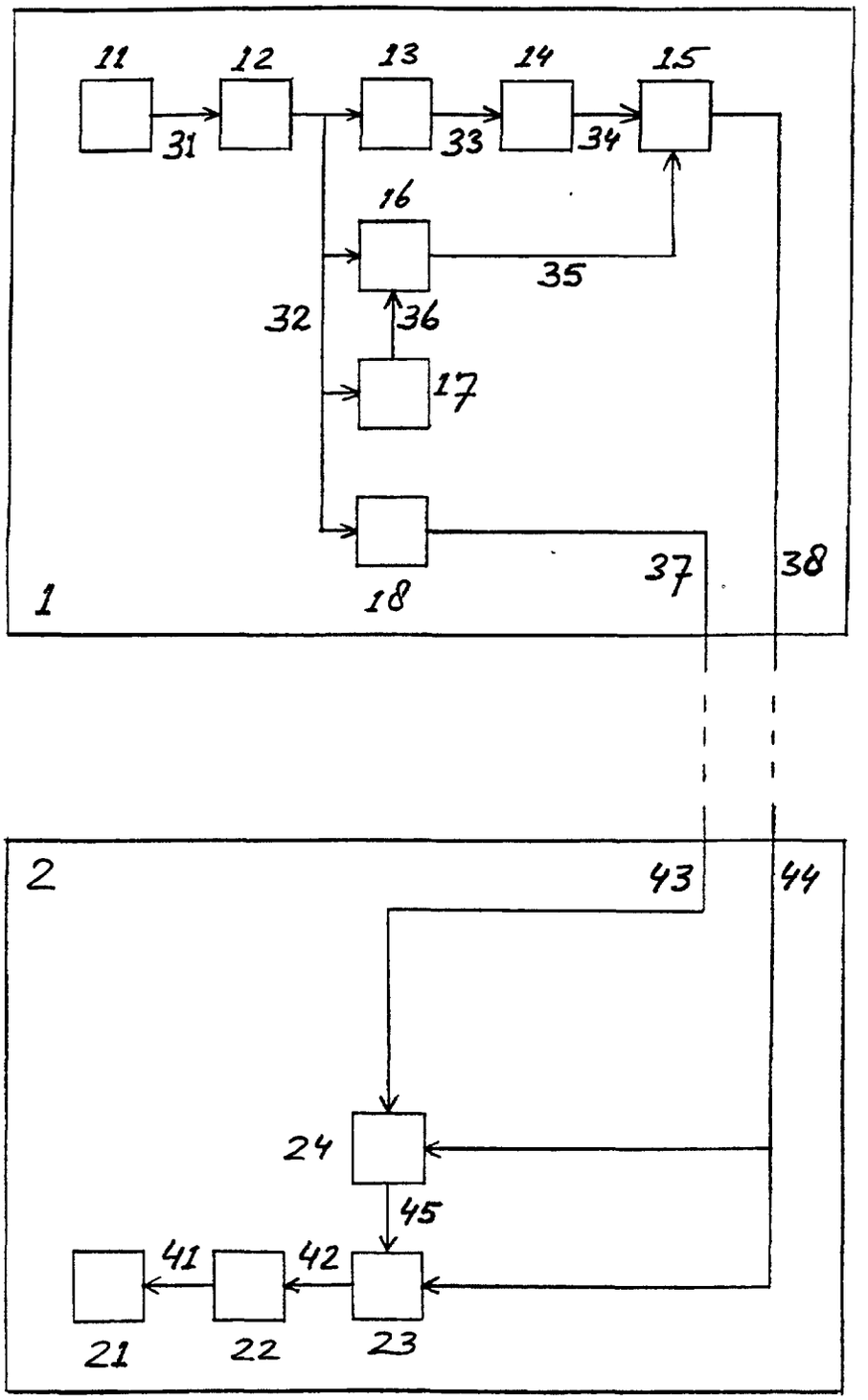
11
12
13
14
15
31
33
34
16
35
32
36
17
18
37
38
1
2
43
44
24
45
41
42
21
22
23

Fig. 1

European Patent Office

EUROPEAN SEARCH REPORT

Application Number
EP 00 44 0335

DOCUMENTS CONSIDERED TO BE RELEVANT

Category	Citation of document with indication, where appropriate, of relevant passages	Relevant to claim	CLASSIFICATION OF THE APPLICATION (Int.Cl.7)
X Y	US 6 073 094 A (CHANG LU ET AL) 6 June 2000 (2000-06-06) * abstract * * column 2, line 59 - column 3, line 4 * * column 3, line 29 - column 3, line 46 * * column 4, line 11 - column 4, line 34 * * claim 14 *	1-3,5-7, 9,10 4,8	G10L19/00
X Y	US 6 119 086 A (ITTYCHERIAH ABRAHAM ET AL) 12 September 2000 (2000-09-12) * column 3, line 23 - line 27 * * abstract *	1-3,5-7, 9,10 4,8	
Y	EP 0 423 800 A (MATSUSHITA ELECTRIC IND CO LTD) 24 April 1991 (1991-04-24) * abstract *	4,8	
			TECHNICAL FIELDS SEARCHED (Int.Cl.7) G10L

The present search report has been drawn up for all claims

Place of search	Date of completion of the search	Examiner
MUNICH	13 September 2001	Bourdier, R

CATEGORY OF CITED DOCUMENTS

X : particularly relevant if taken alone
Y : particularly relevant if combined with another document of the same category
A : technological background
O : non-written disclosure
P : intermediate document

T : theory or principle underlying the invention
E : earlier patent document, but published on, or after the filing date
D : document cited in the application
L : document cited for other reasons

& : member of the same patent family, corresponding document

EPO FORM 1503 03.82 (P04C01)

ANNEX TO THE EUROPEAN SEARCH REPORT ON EUROPEAN PATENT APPLICATION NO. EP 00 44 0335

This annex lists the patent family members relating to the patent documents cited in the above-mentioned European search report. The members are as contained in the European Patent Office EDP file on
The European Patent Office is in no way liable for these particulars which are merely given for the purpose of information.

13-09-2001

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 6073094 A	06-06-2000	AU 3756599 A WO 9963519 A	20-12-1999 09-12-1999
US 6119086 A	12-09-2000	NONE	
EP 0423800 A	24-04-1991	JP 3132797 A DE 69016568 D	06-06-1991 16-03-1995

EPO FORM P0459

For more details about this annex : see Official Journal of the European Patent Office, No. 12/82