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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
18 December 2003 (18.12.2003)

PCT

(10) International Publication Number
WO 03/105124 A1

(51) International Patent Classification[7]: G10L 15/00, H04R 1/40

(21) International Application Number: PCT/US03/18189

(22) International Filing Date: 9 June 2003 (09.06.2003)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
60/388,123 11 June 2002 (11.06.2002) US

(71) Applicant: **SONY ELECTRONICS INC.** [US/US]; 1 Sony Drive, Park Ridge, NJ 07656 (US).

(72) Inventors: **ALMSTRAND, Lars C.**; 3530 Caminito El Rincon #35, San Diego, CA 92130 (US). **KONOPKA, Courtney**; 2399 Jefferson Street #20, Carlsbad, CA 92008 (US).

(74) Agent: **ROGITZ, John L.**; Rogitz & Associates, 750 B Street, Suite 3120, San Diego CA 92101 (US).

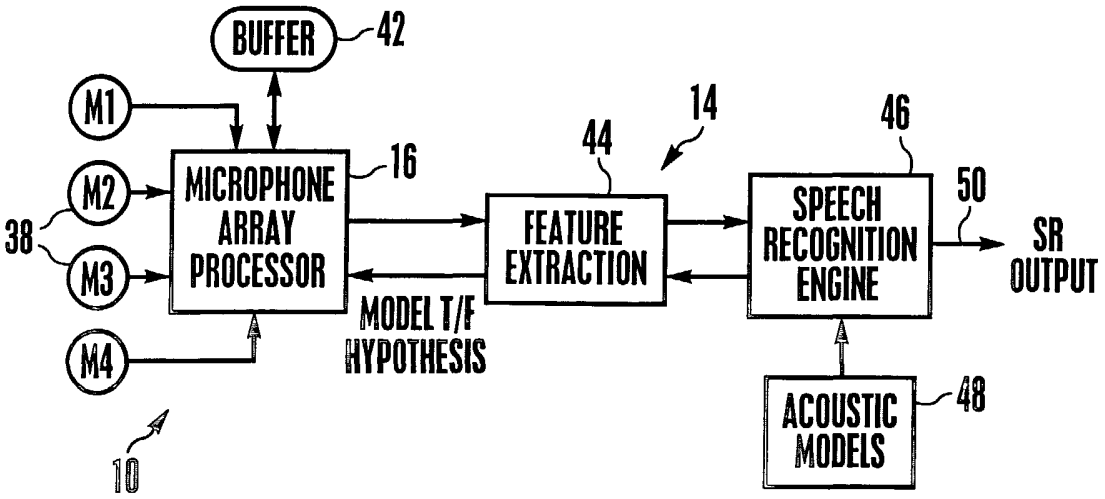
(81) Designated States *(national)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NI, NO, NZ, OM, PH, PL, PT, RO, RU, SC, SD, SE, SG, SK, SL, TJ, TM, TN, TR, TT, TZ, UA, UG, UZ, VC, 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, ES, FI, FR, GB, GR, HU, IE, IT, LU, MC, NL, PT, RO, SE, SI, SK, TR), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:
— *with international search report*
— *before the expiration of the time limit for amending the claims and to be republished in the event of receipt of amendments*

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.

(54) Title: MICROPHONE ARRAY WITH TIME-FREQUENCY SOURCE DISCRIMINATION



(57) Abstract: A 3D microphone array (10) is provided fro, e.g., sending sound to a speech recognition (SR) engine (46). To increase SNR while minimizing the compulational load on the SR engine (46), the array processor (40) compares received time-frequency profiles from various sources (18, 26, 30, 34) of sound to a model hypothesis provided by the SR engine (46) (46), and sends on to the SR engine (46) (46) only those profiles that are similar to the model hypothesis. If desired, sources of sound can also be discriminated against on the basis of energy level and spatial location within a room.

WO 03/105124 A1

MICROPHONE ARRAY WITH TIME-FREQUENCY SOURCE DISCRIMINATION

I. Field of the Invention

The present invention relates generally to microphone arrays.

II. Background of the Invention

Speech recognition devices are well-known. They are used primarily for applications such as word processing, wherein a person speaks into a microphone and phonemes (parts of speech) are recognized by their acoustic patterns and then converted to binary representations and combined into words. In this way, speech can be directly converted into an electronic text file.

For speech recognition devices to work best, the acoustic signals that are processed must have a fairly high signal-to-noise ratio (SNR). The SNR of speech spoken directly into a microphone that is held in front of the speaker's mouth is relatively high, to suit this requirement. On the other hand, the SNR of acoustic signals received on, e.g., an open microphone or speakerphone that is located, for instance, on a table in the middle of a room is generally lower. Consequently, while such microphones are convenient from the standpoint of allowing a person to roam about a room while talking and doing other tasks, the present invention recognizes that to use such microphones as input devices for speech recognition engines, the SNR must be augmented over what might otherwise be afforded.

As understood herein, a microphone array (i.e., several microphones arranged in an array and coupled to a central microphone processor) can be used as a speaker phone, with array microphones providing a directional capability such that the array processor can "form

a beam" (i.e., focus) on sound from specific directions while ignoring sound from other directions. In this way, the SNR of the sound that is processed advantageously is increased.

The present invention further understands, however, that beamforming itself can be processor-intensive. Moreover, it is recognized herein that many sources of sound can be present in a room, which would require many beams to be formed and thus require the speech recognition engine to discriminate the sought-after beam from the various other beams that are received from the microphone array. Accordingly, the present invention provides the solutions disclosed herein.

SUMMARY OF THE INVENTION

A microphone array system includes plural microphones and an array processor that receives signals from the microphones. The array processor executes logic which includes receiving a model time-frequency acoustic hypothesis. Based on the model time-frequency acoustic hypothesis, the array processor selectively outputs signals that represent acoustic sources to a client component, such as but not limited to a speech recognition engine.

In a preferred embodiment, the logic executed by the array processor can include selectively outputting signals to the client component based on acoustic energy levels received from the acoustic sources. Moreover, the logic executed by the array processor may further include selectively outputting signals to the client component based on whether the sources are in a predetermined space. A buffer can be provided to store signals while the array processor executes the logic, with data in the buffer being selectively sent to the client component.

When the client component is a speech recognition device, it can include a feature extraction component that receives signals from the array processor and that sends signals to the speech recognition engine. In this exemplary embodiment, the model time-frequency acoustic hypothesis is generated by sending a signal from the speech recognition engine to the feature extraction component and generating the hypothesis at the feature extraction component, prior to providing a time-frequency representation of the hypothesis to the array processor. As set forth further below, the model hypothesis may represent at least one phoneme.

In another aspect, a method for alleviating processing load on a speech recognition system by screening signals from acoustic sources in a space includes comparing at least one signal from at least one acoustic source in the space to at least one acoustic model. Based at least in part on the comparing act, the signal is selectively sent to the speech recognition system.

In yet another aspect, a device is disclosed that embodies means for processing acoustic signals received from acoustic sources in a volume. The device includes means for comparing signals from sources in the volume to a predefined time-frequency hypothesis. Means send signals to the speech recognition system, responsive to the means for comparing.

The details of the present invention, both as to its structure and operation, can best be understood in reference to the accompanying drawings, in which like reference numerals refer to like parts, and in which:

BRIEF DESCRIPTION OF THE DRAWINGS

Figure 1 is a schematic diagram of a space in which the present microphone array is disposed, showing schematic representations of acoustic signals from various sources;

Figure 2 is a schematic diagram of one preferred architecture; and

Figure 3 is a logic flow chart of the present invention.

DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENT

Referring initially to Figure 1, a microphone array is shown, generally designated 10, that can receive sound from a space 12 and output electrical signals representing the sound to a client component 14. In preferred, non-limiting embodiments the array 10 is a three dimensional array and the client component 14 includes a speech recognition (SR) device, such as speech recognition software with prescreening components described below, that can provide one or model hypotheses (designated at 16) to the array 10. In the preferred, non-limiting embodiment, the model hypothesis 16 is a time (on the x-axis)-frequency (on the y-axis) Cartesian profile, although other types of hypotheses, including energy profiles, can be used. It is to be understood that instead of being a three-dimensional array (i.e., an array with at least four microphones, any three of which establish a plane and the fourth being distanced from the plane), the array 10 can also be a two-dimensional array. It is to be further understood that client components other than SR devices can be used, e.g., the client component 14 might be a sound speaker that is to play only predetermined sounds, e.g., bird chirps, that conform to the model hypothesis 16.

In the non-limiting illustrative environment shown in Figure 1, various sources of sound are shown in the space 12, along with graphic representations of the sound they emit. Beginning in the lower right corner, a TV 18 might produce sound at least a portion of which establishes a time-frequency (T/F) profile 20 that, as shown, includes upside-down semicircles separated by a dot. Also, sound can emanate from a door 22 that has an acoustic energy profile 24, and from a window 26 that has, as shown at 28, a relatively lower acoustic energy profile than the door 22.

Also, a person 30 might be speaking in the room, with at least a portion of the speech establishing a T/F profile 32 that closely resembles the model hypothesis shown at 16. Finally, a radio 34 can play sound having at least in part a T/F profile 36 that, like the exemplary model hypothesis shown, is characterized by two curves that extend up and to the right and that are separated by a dot, but one that, unlike the model hypothesis shown for illustration, is not characterized by a dogleg in each curve.

One exemplary, non-limiting architecture is shown in Figure 2. As shown, the array 10 can include plural microphones 38 that receive acoustic energy and output electrical signals representative thereof to an array processor system 40. The processor system 40 can include a digital processor proper as well as necessary digitizing components known in the art.

The processor 40 can further access a data buffer 42 to store digitized signals in the buffer 42 pending the results of the logic disclosed below, prior to sending the signals on to the client component 14. In the preferred embodiment shown in Figure 2, the client component 14 can be a speech recognition (SR) device that includes a feature extraction component 44 for extracting key features of signals received from the array processor 40, a

speech recognition engine 46 that receives the output of the feature extraction component 44, and acoustic models 48 that are used by the SR engine 46 in accordance with means known in the art to transform electrical signals representing sound into electronic text (or other) tokens for output thereof as indicated at 50. As also shown, the model T/F hypotheses mentioned above and discussed further below can be sent from the SR engine 46 by way of the feature extraction component 44 to the array processor 40. The SR engine 46 may access a spelling dictionary and hidden Markov models in accordance with SR operating principles known in the art.

The logic of the array processor 40 can be seen in reference to Figure 3, it being understood that the below-described logic may be executed in whole or in part by the client component 14 if desired. Commencing at block 52, the model hypothesis is established. In one preferred, non-limiting embodiment, the model hypothesis might represent a predetermined acoustic temporal pattern such as an acoustic sub-word or series of sub-words, such as a signalling word like "Mona" that can be programmed into a word spotter implemented as a standalone component or integrated component in the processor 40 or outside of the processor 40. To establish the model, the SR engine 46 can cooperate with the feature extraction component 44 to transform electronic symbols representative of, e.g., "Mona", to the T/F graph shown in the model hypothesis box 16 of Figure 1, essentially by reverse SR. Or, the model hypothesis can be any other T/F graph as desired by the user, e.g., a graph representing bird chirps. At block 54, the model hypothesis is sent to the array processor 40/word spotter.

Block 56 indicates that if desired, spatial localization can be enabled and used to pre-screen sounds received by the array 10. Specifically, since the array 10 is a multi-dimensional array, a predetermined space from which sounds will be subsequently processed can be defined during a calibration process, with sounds emanating from locations outside the predetermined space being attenuated during subsequent processing. The predetermined space can be defined by means well known in the art, e.g., by using geometric triangulation to correlate differences among the microphones 38 in the times of reception of a sound wave to the spatial boundaries of the desired volume. Thus, for instance, the boundaries of the entire space 12 shown in Figure 1 can be predetermined to be the space of consideration, with sound emanating from points outside the space 12 being attenuated, or only a portion of the space 12 might be predetermined to be the space of interest, with sounds emanating from outside the portion being disregarded.

To define the space, a calibration process can be used. As one non-limiting example, a beeper that transmits a sine wave can be located at various points along the desired boundary of the space and activated, with the system being set in a calibration mode such that the system receives the beeps, triangulates the position of the source beeper to find the position, and then stores the positions as a map of the space boundary. Moreover, when a known sine wave is used, any distortions, amplifications, or attenuations can be noted from the various locations and either the user informed not to stand at distorting locations or the system adjusted as appropriate to cancel out the distortions, e.g., by amplifying all subsequent sounds from a location at which the beeper signal experienced attenuation during calibration.

As indicated at block 56, for sounds emanating from inside the predetermined space of interest, such sounds are temporarily buffered while being passed to the next state of the logic at block 58. This is an optional state that further pre-screens sounds on the basis of energy level, with signals exhibiting acceptable levels being passed to the T/F discrimination state at block 60. To illustrate, the relatively low energy level sounds (represented by the graph 28 in Figure 1) from the window 26 might be screened out from further processing at this point in the logic, whereas higher-level sounds such as that of the door graph 24, person profile 32, TV profile 20, and radio profile 36 might be passed on.

If desired, sounds that pass the logic at blocks 58 and 60 can be sent to a word spotter that is programmed to recognize just a few unique signalling words (such as "Mona") and which functions in accordance with principles known in the art to pass on any candidate signalling words to the speech recognition engine, such that the system can focus on the location of the source of the signalling word. The user can update the signalling words used by the word spotter. Or, dynamic time warping (DTW) can be used at block 61 to modify a received signal to match a hypothesis entry.

At block 62 the sound profiles that have satisfied the above spatial pre-screening conditions (if enabled), the energy level pre-screening conditions (if enabled), and word spotting/DTW conditions (if enabled) are compared to the model hypothes(es). Only sounds bearing a T/F profile that is sufficiently similar to the model hypothes(es) are passed, at block 62, from the buffer 42 to the client component 14 (which can include, e.g., the SR system shown in Figure 2) for, e.g., speech recognition of the signals at block 64. The comparison between the signal curves from the various sources and the curve(s) of the model hypothes(es)

can be made in accordance with signal comparison principles known in the art, e.g., it can be based on a least-squares fit, point by point, or on some other signal comparison paradigm.

Again using the example profiles in Figure 1, at block 62 the TV profile 20 and door profile 24 might be filtered out and not passed on to the client component 14. In contrast, the person profile 32 and radio profile 36 might both exhibit sufficient resemblance to the model hypothesis to warrant sending signals from these sources on to the client component 14. It will be readily appreciated that by eliminating one or more sources of sound in this way, the array processor 40 relieves the client component 14 of significant processing load.

What constitutes sufficient similarity between a source T/F profile and the model hypothes(es) varies from application to application, and can be done by raw curve analysis that accounts for the processing capability of the client component 14. Essentially, the risk of losing what might be valuable data is balanced on an application-by-application basis against the increased performance afforded to the particular client component 14 by the above-described pre-discrimination of signals, to establish what constitutes sufficient similarity.

With the above disclosure in mind, it may now be appreciated that the present invention provides a multidimensional microphone array that is tightly coupled to its client component and that can pre-screen acoustic sources in parallel with the processing being undertaken by the client component. Moreover, it may now be appreciated that the array 10 described herein is dynamic, in that the model hypothes(es) can be changed as desired to change what T/F profiles are passed on to, e.g., the SR engine 46 shown in Figure 2.

It is to be understood that while for convenience the above logic is described in terms of logical states such as might be assumed by a state machine, the logic can also be thought of as a series of decision steps, wherein the logic flows from one decision step to the next when certain conditions test positive, e.g., spatial location, energy level, and finally T/F profile similarity.

The logic may be executed by a processor or processors within the present system as a series of computer-executable instructions. The instructions may be contained on a data storage device with a computer readable medium, such as a computer diskette having a computer usable medium with code elements stored thereon. Or, the instructions may be stored on random access memory (RAM) of the computer, or on conventional hard disk drive, electronic read-only memory, optical storage device, or other appropriate data storage device.

Indeed, the flow charts herein illustrate the structure of the logic of the present invention as might be embodied in computer program software. Those skilled in the art will appreciate that the flow charts illustrate the structures of computer program code elements that may function according to this invention. Manifestly, the invention can be practiced in its essential embodiment by a machine component that renders the program code elements in a form that instructs a digital processing apparatus (that is, a computer) to perform a sequence of function steps corresponding to those shown.

While the particular MICROPHONE ARRAY WITH TIME-FREQUENCY SOURCE DISCRIMINATION as herein shown and described in detail is fully capable of attaining the above-described objects of the invention, it is to be understood that it is the presently preferred embodiment of the present invention and is thus representative of the subject matter

which is broadly contemplated by the present invention, that the scope of the present invention fully encompasses other embodiments which may become obvious to those skilled in the art, and that the scope of the present invention is accordingly to be limited by nothing other than the appended claims, in which reference to an element in the singular is not intended to mean "one and only one" unless explicitly so stated, but rather "one or more".

WE CLAIM:

WHAT IS CLAIMED IS:

1. A microphone array system, comprising:

plural microphones (38); and

at least one array processor (40) receiving signals from the microphones (38), the array processor (40) executing logic including:

receiving at least one model time-frequency acoustic hypothesis; and

based at least in part on the model time-frequency acoustic hypothesis, selectively outputting signals representing at least one acoustic source to at least one client component (14).

2. The system of Claim 1, wherein the client component includes at least one speech recognition engine (46), and the system further comprises the engine (46).

3. The system of Claim 1, comprising at least four microphones (38) establishing a three dimensional array.

4. The system of Claim 1, wherein the logic executed by the array processor (40) further includes selectively outputting signals to the client component (14) based at least in part on at least one acoustic energy level received from at least one acoustic source.

5. The system of Claim 1, wherein the logic executed by the array processor (40) further includes selectively outputting signals to the client component (14) based at least in part on at least one spatial location of at least one acoustic source.

6. The system of Claim 1, further comprising at least one buffer (42) storing signals while the array processor (40) executes the logic, data in the buffer (42) being selectively sent to the client component (14).

7. The system of Claim 2, further comprising at least one feature extraction component (44) receiving signals from the array processor (40) and sending signals to the speech recognition engine (46).

8. The system of Claim 7, wherein the model time-frequency acoustic hypothesis is generated by sending at least one signal from the speech recognition engine (46) to the feature extraction component (44) and generating a time-frequency representation of the hypothesis at the feature extraction component (44), prior to providing the hypothesis to the array processor (40).

9. The system of Claim 8, wherein the hypothesis represents at least one acoustic temporal pattern.

10. The system of Claim 1, wherein the hypothesis represents at least one acoustic temporal pattern.

11. The system of Claim 1, further comprising the client component (14), wherein the client component (14) includes at least one audio speaker.

12. A method for alleviating processing load on a speech recognition system by screening signals from acoustic sources in a space (12), comprising:

comparing at least one signal from at least one acoustic source (18, 26, 30, 34) in the space (12) to at least one acoustic model; and

based at least in part on the comparing act, selectively sending the signal to the speech recognition system (46).

13. The method of Claim 12, wherein the acoustic model is at least one time-frequency hypothesis.

14. The method of Claim 12, wherein the acoustic model is at least one acoustic energy level.

15. The method of Claim 12, wherein the space (12) is predefined, and the acoustic model is at least whether a source is located in the space (12).

16. The method of Claim 12, comprising receiving signals from acoustic sources at a multidimensional microphone array (10) and undertaking the comparing and sending acts at the array (10).

17. The method of Claim 16, wherein the array (10) is a three dimensional array.

18. The method of Claim 13, wherein the time-frequency hypothesis is received by the array (10) from the speech recognition system (46).

19. A device embodying means for processing acoustic signals received from at least one source (18, 26, 30, 34) in at least one volume (12), comprising:

means for comparing signals from sources in the volume (12) to at least one time-frequency hypothesis; and

means, responsive to the means for comparing, for sending signals to the speech recognition system (46).

20. The device of Claim 19, further comprising:

means for defining the volume; and

means for invoking the comparing means only in response to acoustic signals received from sources within the volume.

21. The device of Claim 19, further comprising:

means for invoking the comparing means only in response to acoustic signals having at least a predetermined energy.

22. The device of Claim 19, further comprising the speech recognition system.

23. The device of Claim 22, wherein the speech recognition system (46) comprises means for defining the time-frequency hypothesis.

24. The device of Claim 19, wherein the means for comparing is executed by a three dimensional microphone array processor (40).

25. The device of Claim 19, wherein the means for comparing and sending are embodied in software.

26. The device of Claim 19, wherein the means for comparing and sending are embodied in hardware.

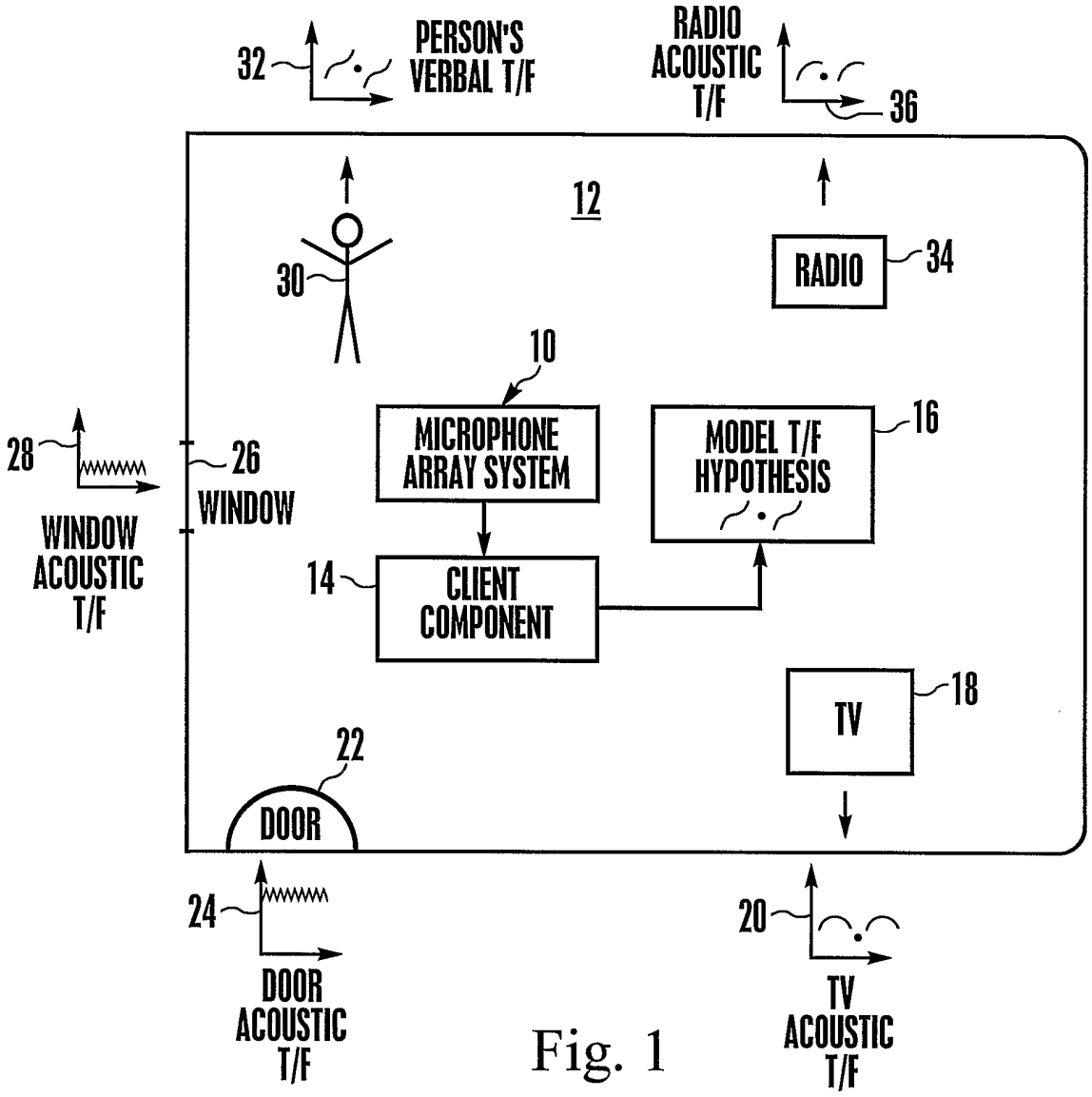
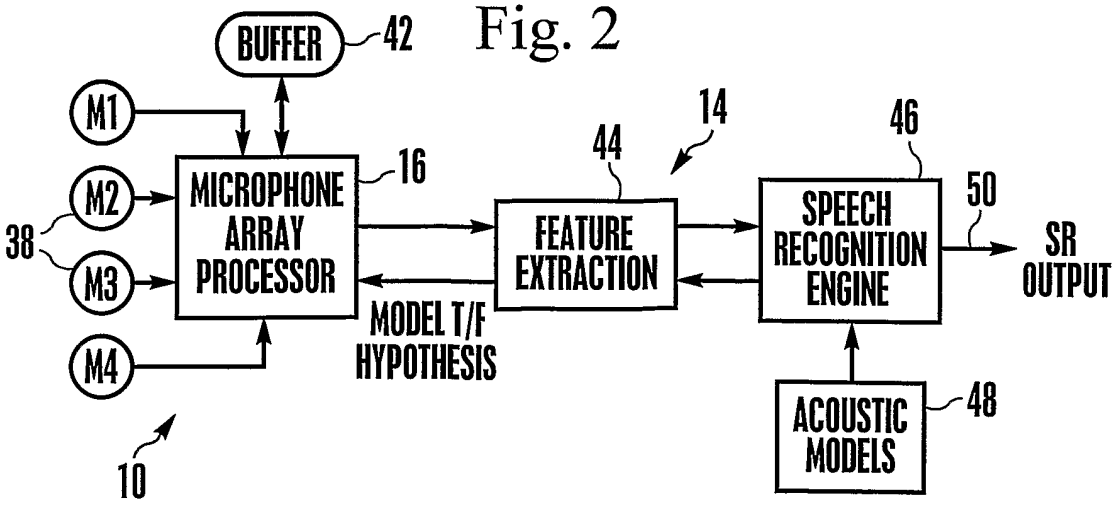

32
PERSON'S VERBAL T/F
RADIO ACOUSTIC T/F
36
12
30
RADIO
34
10
MICROPHONE ARRAY SYSTEM
MODEL T/F HYPOTHESIS
16
28
26
WINDOW
WINDOW ACOUSTIC T/F
14
CLIENT COMPONENT
TV
18
22
DOOR
24
DOOR ACOUSTIC T/F
20
TV ACOUSTIC T/F

Fig. 1

Fig. 2


BUFFER
42
M1
M2
M3
M4
38
MICROPHONE ARRAY PROCESSOR
16
MODEL T/F HYPOTHESIS
44
14
FEATURE EXTRACTION
46
SPEECH RECOGNITION ENGINE
50
SR OUTPUT
ACOUSTIC MODELS
48
10

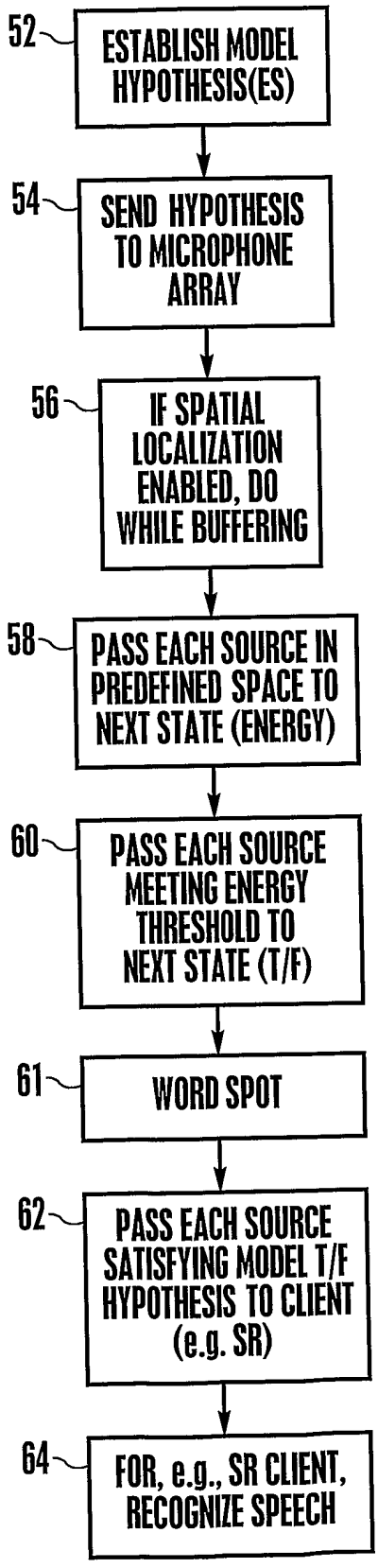

52
ESTABLISH MODEL HYPOTHESIS(ES)
54
SEND HYPOTHESIS TO MICROPHONE ARRAY
56
IF SPATIAL LOCALIZATION ENABLED, DO WHILE BUFFERING
58
PASS EACH SOURCE IN PREDEFINED SPACE TO NEXT STATE (ENERGY)
60
PASS EACH SOURCE MEETING ENERGY THRESHOLD TO NEXT STATE (T/F)
61
WORD SPOT
62
PASS EACH SOURCE SATISFYING MODEL T/F HYPOTHESIS TO CLIENT (e.g. SR)
64
FOR, e.g., SR CLIENT, RECOGNIZE SPEECH

Fig. 3

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US03/18189

A. CLASSIFICATION OF SUBJECT MATTER

IPC(7) : G10L 15/00; H04R 1/40
US CL : 704/231; 367/124; 381/92

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
U.S. : 704/231, 233, 251, 255; 367/124; 381/92

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)
Please See Continuation Sheet

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category *	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X --- Y	US 6,219,645 B1 (BYERS) 17 April 2001 (17.04.2001), column 5, lines 20 to 51; column 6, line 50 to column 7, line 26; column 7, line 27 to column 8, line 47: Figures 1, 2, and 4.	1-6,10,12-17,19-26 ---------- 7-9,11,18
Y	US 5,737,485 A (FLANAGAN et al.) 07 April 1998 (07.04.1998), column 3, line 46 to column 4, line 45: Figure 1.	7-9, 18
Y	US 6,222,927 B1 (FENG et al.) 24 April 2001 (24.04.2001), column 4, lines 38 to 51; column 9, lines 54 to 67: Figures 1 and 4B	11
A	US 5,828,997 A (DURLACH et al.) 27 October 1998 (27.10.1998), column 1, line 65 to column 2, line 19.	1-26
A	US 5,353,376 A (OH et al.) 04 October 1994 (04.10.1994), column 5, line 59 to column 6, line 14.	1-26

☐ Further documents are listed in the continuation of Box C. ☐ See patent family annex.

* Special categories of cited documents:

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
23 September 2003 (23.09.2003)	08 OCT 2003
Name and mailing address of the ISA/US Mail Stop PCT, Attn: ISA/US Commissioner for Patents P.O. Box 1450 Alexandria, Virginia 22313-1450 Facsimile No. (703)305-3230	Authorized officer Richemond Dorvil Telephone No. (703) 305-4700

Form PCT/ISA/210 (second sheet) (July 1998)

Continuation of B. FIELDS SEARCHED Item 3:
EAST search keywords: speech adj recognition and microphone adj array; three adj dimensional and microphone adj array; and feature; and buffer; and speaker