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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
14 November 2013 (14.11.2013)

WIPO | PCT

(10) International Publication Number
WO 2013/169618 A1

(51) International Patent Classification:
G10L 25/48 (2013.01) *G01S 3/80* (2006.01) *G10L 17/00* (2013.01) *G10L 25/06* (2013.01) *H04R 1/40* (2006.01)

(21) International Application Number: PCT/US2013/039624

(22) International Filing Date: 6 May 2013 (06.05.2013)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
61/645,818 11 May 2012 (11.05.2012) US
13/674,690 12 November 2012 (12.11.2012) US

(71) Applicant: **QUALCOMM INCORPORATED** [US/US]; Attn: International IP Administration, 5775 Morehouse Drive, San Diego, California 92121-1714 (US).

(72) Inventors: **KIM, Lae-Hoon**; 5775 Morehouse Drive, San Diego, California 92121-1714 (US). **SHIN, Jongwon**; 123 Cheomdan-gwagiro, Buk-gu, Gwangju 500-712 (KR). **VISSER, Erik**; 5775 Morehouse Drive, San Diego, California 92121-1714 (US).

(74) Agent: **TOLER, Jeffrey G.**; 8500 Bluffstone Cove, Suite A201, Austin, Texas 78759 (US).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) Designated States *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ,

[Continued on next page]

(54) Title: AUDIO USER INTERACTION RECOGNITION AND CONTEXT REFINEMENT

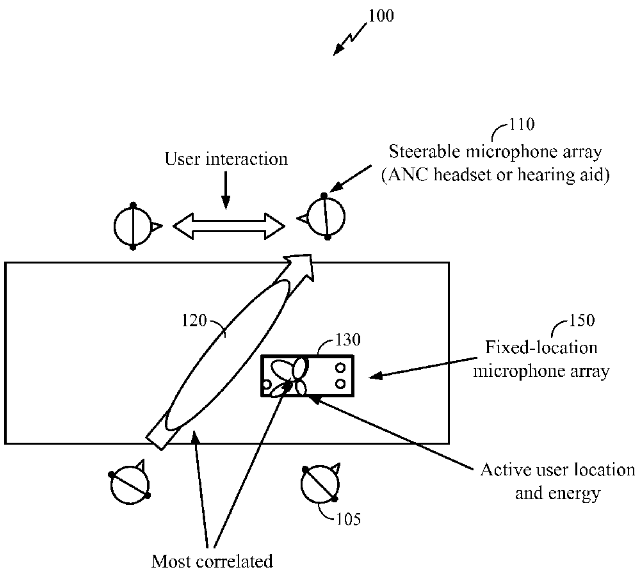


FIG. 1

(57) **Abstract**: A system which tracks a social interaction between a plurality of participants, includes a fixed beamformer that is adapted to output a first spatially filtered output and configured to receive a plurality of second spatially filtered outputs from a plurality of steerable beamformers. Each steerable beamformer outputs a respective one of the second spatially filtered outputs associated with a different one of the participants. The system also includes a processor capable of determining a similarity between the first spatially filtered output and each of the second spatially filtered outputs. The processor determines the social interaction between the participants based on the similarity between the first spatially filtered output and each of the second spatially filtered outputs

WO 2013/169618 A1

TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Declarations under Rule 4.17:

— *as to applicant's entitlement to apply for and be granted a patent (Rule 4.17(ii))*

— *as to the applicant's entitlement to claim the priority of the earlier application (Rule 4.17(iii))*

Published:

— *with international search report (Art. 21(3))*

AUDIO USER INTERACTION RECOGNITION AND CONTEXT REFINEMENT

CROSS-REFERENCE TO RELATED APPLICATIONS

[0001] This application claims priority under the benefit of 35 U.S.C. § 119(e) to U.S. Provisional Patent Application No. 61/645,818, filed May 11, 2012 and entitled "AUDIO USER INTERACTION RECOGNITION AND CONTEXT REFINEMENT," and claims priority to U.S. Non-Provisional Patent Application No. 13/674,690 filed on November 12, 2012 and entitled "AUDIO USER INTERACTION RECOGNITION AND CONTEXT REFINEMENT," the contents of which are expressly incorporated herein by reference.

BACKGROUND

[0002] A substantial amount of useful information can be derived from determining the direction a user is looking at different points in time, and this information can be used to enhance the user's interaction with a variety of computational systems. Therefore, it is not surprising that a vast amount of gaze tracking research using a vision based approach (i.e., tracking the eyes using any of several various means) has already been undertaken. However, understanding a user's gazing direction only gives semantic information on one dimension of the user's interest and does not take into account contextual information that is mostly given by speech. In other words, the combination of gaze tracking coupled with speech tracking would provide richer and more meaningful information in a variety of different user applications.

SUMMARY

[0003] Contextual information (that is, non-visual information that is being sent or received by a user) is determined using an audio based approach. Audio user interaction on the receiving side may be enhanced by steering audio beams toward a specific person or a specific sound source. The techniques described herein may therefore allow a user to more clearly understand the context of a conversation, for example. To achieve these benefits, inputs from one or more steerable microphone arrays and inputs from a fixed microphone array may be used to determine who a person is looking at or what a person is paying attention to relative to who is speaking where audio-based contextual information (or even visual-based semantic information) is being presented.

[0004] For various implementations, two different types of microphone array devices (MADs) are used. The first type of MAD is a steerable microphone array (also referred to herein as a steerable array) which is worn by a user in a known orientation with regard to the user's eyes, and multiple users may each wear a steerable array. The second type of MAD is a fixed-location microphone array (also referred to herein as a fixed array) which is placed in the same acoustic space as the users (one or more of which are using steerable arrays).

[0005] For certain implementations, the steerable microphone array may be part of an active noise control (ANC) headset or hearing aid. There may be multiple steerable arrays, each associated with a different user or speaker (also referred to herein as a participant) in a meeting or group, for example. The fixed microphone array, in such a context, would then be used to separate different people speaking and listening during the group meeting using audio beams corresponding to the direction in which the different people are located relative to the fixed array.

[0006] The correlation or similarity between the audio beams of the separated speakers of fixed array and the outputs of the steerable arrays are evaluated. Correlation is one example of a similarity measure, although any of several similarity measurement or determination techniques may be used.

[0007] In an implementation, the similarity measure between the audio beams of the separated participants of the fixed array and the outputs of steerable arrays may be used to track social interaction between participants, including gazing direction of the participants over time as different participants speak or present audio-based information.

[0008] In an implementation, the similarity measure between the audio beams of the separated participants of the fixed array and the outputs of steerable arrays may be used to zoom in on a targeted participant, for example. This zooming might in turn lead to enhanced noise filtering and amplification when one user (who at that moment is a listener) is gazing at another person who is providing audio-based information (i.e., speaking).

[0009] In an implementation, the similarity measure between the audio beams of the separated participants of the fixed array and the outputs of steerable arrays may be used to adaptively form a better beam for a targeted participant, in effect better determining the physical orientation of each of the users relative to each other.

[0010] This summary is provided to introduce a selection of concepts in a simplified form that are further described below in the detailed description. This summary is not intended to identify key features or essential features of the claimed subject matter, nor is it intended to be used to limit the scope of the claimed subject matter.

BRIEF DESCRIPTION OF THE DRAWINGS

[0011] The foregoing summary, as well as the following detailed description of illustrative embodiments, is better understood when read in conjunction with the appended drawings. For the purpose of illustrating the embodiments, there are shown in the drawings example constructions of the embodiments; however, the embodiments are not limited to the specific methods and instrumentalities disclosed. In the drawings:

[0012] Figure 1 is a diagram of a group of users each wearing a steerable microphone array, along with a fixed microphone array, that may be used to determine contextual information;

[0013] Figure 2 is an operational flow of an implementation of a method of determining user interaction using steerable microphone arrays and a fixed microphone array;

[0014] Figure 3 is an operational flow of another implementation of a method of determining user interaction using steerable microphone arrays and a fixed microphone array;

[0015] Figure 4 is a diagram of an example display that may provide an indication of a user identity and which direction the user is looking;

[0016] Figure 5 is a diagram of a user interface that may be generated and displayed and that indicates various user interactions and meeting data;

[0017] Figure 6 is a diagram of an example display of a user interface that may be generated and displayed (e.g., on a smartphone display) and that indicates various user interactions (e.g., during a meeting);

[0018] Figure 7 is a diagram of an example display that indicates various user interactions with respect to various topics;

[0019] Figure 8 is a diagram of an example display that indicates various user interactions over time;

[0020] Figure 9 is a diagram of another example display that indicates various user interactions over time;

[0021] Figure 10 is an operational flow of an implementation of a method of a measuring similarity using cross-correlation;

[0022] Figure 11 is an operational flow of an implementation of a method of measuring similarity using cross-cumulant;

[0023] Figure 12 is an operational flow of an implementation of a method of measuring similarity using time-domain least squares fit;

[0024] Figure 13 is an operational flow of an implementation of a method of measuring similarity using frequency-domain least squares fit;

[0025] Figure 14 is an operational flow of an implementation of a method of measuring similarity using Itakura-Saito distance;

[0026] Figure 15 is an operational flow of an implementation of a method of measuring similarity using a feature based approach;

[0027] Figure 16 shows an example user interface display;

[0028] Figure 17 shows an exemplary user interface display to show collaborative zooming on the display;

[0029] Figure 18 is an operational flow of an implementation of a method for zooming into a target participant;

[0030] Figure 19 shows an example user interface display with additional candidate look directions;

[0031] Figure 20 is an operational flow of an implementation of a method for adaptively refining beams for a targeted speaker;

[0032] Figure 21 shows a far-field model of plane wave propagation relative to a microphone pair;

[0033] Figure 22 shows multiple microphone pairs in a linear array;

[0034] Figure 23 shows plots of unwrapped phase delay vs. frequency for four different DOAs, and Figure 24 shows plots of wrapped phase delay vs. frequency for the same DOAs;

[0035] Figure 25 shows an example of measured phase delay values and calculated values for two DOA candidates;

[0036] Figure 26 shows a linear array of microphones arranged along the top margin of a television screen;

[0037] Figure 27 shows an example of calculating DOA differences for a frame;

[0038] Figure 28 shows an example of calculating a DOA estimate;

[0039] Figure 29 shows an example of identifying a DOA estimate for each frequency;

[0040] Figure 30 shows an example of using calculated likelihoods to identify a best microphone pair and best DOA candidate for a given frequency;

[0041] Figure 31 shows an example of likelihood calculation;

[0042] Figure 32 shows an example of a speakerphone application;

[0043] Figure 33 shows a mapping of pair-wise DOA estimates to a 360° range in the plane of the microphone array;

[0044] Figures 34 and 35 show an ambiguity in the DOA estimate;

[0045] Figure 36 shows a relation between signs of observed DOAs and quadrants of an x-y plane;

[0046] Figures 37-40 show an example in which the source is located above the plane of the microphones;

[0047] Figure 41 shows an example of microphone pairs along non-orthogonal axes;

[0048] Figure 42 shows an example of use of the array of Figure 41 to obtain a DOA estimate with respect to the orthogonal x and y axes;

[0049] Figures 43 and 44 show examples of pair-wise normalized beamformer/null beamformers (BFNFs) for a two-pair microphone array (e.g., as shown in Figure 45);

[0050] Figure 46 shows an example of a pair-wise normalized minimum variance distortionless response (MVDR) BFNF;

[0051] Figure 47 shows an example of a pair-wise BFNF for frequencies in which the matrix A^HA is not ill-conditioned;

[0052] Figure 48 shows examples of steering vectors; and

[0053] Figure 49 shows a flowchart of an integrated method of source direction estimation as described herein.

DETAILED DESCRIPTION

[0054] Unless expressly limited by its context, the term “signal” is used herein to indicate any of its ordinary meanings, including a state of a memory location (or set of memory locations) as expressed on a wire, bus, or other transmission medium. Unless expressly limited by its context, the term “generating” is used herein to indicate any of its ordinary meanings, such as computing or otherwise producing. Unless expressly limited by its context, the term “calculating” is used herein to indicate any of its ordinary meanings, such as computing, evaluating, estimating, and/or selecting from a plurality of values. Unless expressly limited by its context, the term "obtaining" is used to indicate any of its ordinary meanings, such as calculating, deriving, receiving (e.g., from an external device), and/or retrieving (e.g., from an array of storage elements). Unless expressly limited by its context, the term "selecting" is used to indicate any of its ordinary meanings, such as identifying, indicating, applying, and/or using at least one, and fewer than all, of a set of two or more. Where the term “comprising” is used in the present description and claims, it does not exclude other elements or operations. The term “based on” (as in “A is based on B”) is used to indicate any of its ordinary meanings, including the cases (i) “derived from” (e.g., “B is a precursor of A”), (ii) “based on at least” (e.g., “A is based on at least B”) and, if appropriate in the particular context, (iii) “equal to” (e.g., “A is equal to B” or “A is the same as B”). Similarly, the term “in response to” is used to indicate any of its ordinary meanings, including “in response to at least.”

[0055] References to a “location” of a microphone of a multi-microphone audio sensing device indicate the location of the center of an acoustically sensitive face of the microphone, unless otherwise indicated by the context. The term “channel” is used at times to indicate a signal path and at other times to indicate a signal carried by such a path, according to the particular context. Unless otherwise indicated, the term “series” is used to indicate a sequence of two or more items. The term “logarithm” is used to indicate the base-ten logarithm, although extensions of such an operation to other bases are within the scope of this disclosure. The term “frequency component” is used to indicate one among a set of frequencies or frequency bands of a signal, such as a sample (or “bin”) of a frequency domain representation of the signal (e.g., as produced by a fast Fourier transform) or a subband of the signal (e.g., a Bark scale or mel scale subband).

[0056] Unless indicated otherwise, any disclosure of an operation of an apparatus having a particular feature is also expressly intended to disclose a method having an analogous feature (and vice versa), and any disclosure of an operation of an apparatus according to a particular configuration is also expressly intended to disclose a method according to an analogous configuration (and vice versa). The term “configuration” may be used in reference to a method, apparatus, and/or system as indicated by its particular context. The terms “method,” “process,” “procedure,” and “technique” are used generically and interchangeably unless otherwise indicated by the particular context. The terms “apparatus” and “device” are also used generically and interchangeably unless otherwise indicated by the particular context. The terms “element” and “module” are typically used to indicate a portion of a greater configuration. Unless expressly limited by its context, the term “system” is used herein to indicate any of its ordinary meanings, including “a group of elements that interact to serve a common purpose.”

[0057] Any incorporation by reference of a portion of a document shall also be understood to incorporate definitions of terms or variables that are referenced within the portion, where such definitions appear elsewhere in the document, as well as any figures referenced in the incorporated portion. Unless initially introduced by a definite article, an ordinal term (e.g., “first,” “second,” “third,” etc.) used to modify a claim element does not by itself indicate any priority or order of the claim element with respect to another, but rather merely distinguishes the claim element from another claim element having a same name (but for use of the ordinal term). Unless expressly limited by its context, each of the terms “plurality” and “set” is used herein to indicate an integer quantity that is greater than one.

[0058] A combination visual- and hearing-based approach is described herein to enable a user to steer towards a person (or a sound source) in order to more clearly understand the audio-based information being presented at that moment (e.g., the context of conversation and/or the identity of the sound source) using sound sensors and a variety of position-based calculations and resulting interaction enhancements.

[0059] For example, the correlation or similarity between the audio beams of the separated speakers of the fixed array and the outputs of steerable arrays may be used to track social interaction between speakers. Correlation is just one example of a similarity measure, and any similarity measurement or determination technique may be used.

[0060] More particularly, a social interaction or social networking analysis of a group of users (also referred to herein as speakers or participants) may be performed and displayed using a connection graph generated responsive to the correlation or other similarity measure between the audio beams of the separated speakers of the fixed array and the output of each steerable array respectively associated with each user of the group. Thus, for example, automatic social network analysis may be performed in a group meeting of participants, using a connection graph among the meeting participants, to derive useful information regarding who was actively engaged in the presentation or more generally the effectiveness of the presentation in holding the attention of the users.

[0061] Figure 1 is a diagram 100 of a group of users each wearing a steerable microphone array 110, along with a fixed-location microphone array 150 in the same space (e.g., room) as the users, which may be used to determine contextual information. As shown in Figure 1, each user 105 of a group of users in a room (or other defined space) wears a steerable microphone array (e.g., as a headset that may include the ability to perform adaptive noise control (ANC)), and a fixed-location microphone array 150 is located in the room (e.g., on a table, in a phone, etc.). The fixed-location microphone array 150 may be part of an electronic device such as a video game platform, tablet, notebook, or smartphone, for example, or may be a standalone device or implementation. Alternatively or additionally, the fixed-location microphone array 150 may comprise a distributed microphone array (i.e., distributed microphones).

[0062] A user 105 wearing the headset may generate a fixed beam-pattern 120 from his steerable (e.g., wearable) microphone array which is pointed in the user's physical visual (or "look") direction. If the user turns his head, then the user's look direction of the beam-pattern is also changed. The active speaker's location may be determined using the fixed microphone array. By correlating, or otherwise determining the similarity of, beamformed output (or any type of spatially filtered output) from the steerable microphone array with the fixed microphone array outputs corresponding to each active speaker, the identification may be determined of the person that a user is looking at (e.g., paying attention to, listening to, etc.). Each headset may be have processor that is in communication (e.g., via a wireless communications link) with a main processor (e.g., in a centralized local or remote computing device) to analyze correlations or similarities of beams between the headsets and/or the fixed arrays.

[0063] In other words, fixed beam patterns at any moment in time may be formed based on a user's physical look direction which can be correlated with the fixed microphone array outputs, thereby providing a visual indication, via a connection graph 130 (e.g., displayed on a display of any type of computing device, such as a handset, a laptop, a tablet, a computer, a netbook, or a mobile computing device), of the social interaction of the targeted users. Thus, by correlating a beamformed output from the steerable microphone array with the fixed microphone array outputs, corresponding to each active speaking user, tracking of a social interaction or network analysis may be performed and displayed. Moreover, by checking the similarity between beamformed output from the look-direction-steerable microphone array and the location-fixed microphone array outputs corresponding to each active speaker, the person that a user is looking at or paying attention to can be identified and zoomed into.

[0064] Figure 2 is an operational flow of an implementation of a method 200 of determining user interaction using steerable microphone arrays and a fixed microphone array. At 210, the steerable microphone arrays and the fixed microphone array each receive sound at roughly the same time (although small variations can be detected and used to calculate relative positions of the user). At 220, a spatially filtered output, such as a beamformed output, is generated by each of the steerable microphone arrays and the fixed microphone array. At 230, the spatially filtered output of each steerable microphone array is compared with the spatially filtered output of the fixed microphone array. Any known technique for determining similarity or correlation may be used. At 240, the similarity or correlation information obtained from 230 may be used to determine and/or display user interaction information, as described further herein.

[0065] Figure 3 is an operational flow of another implementation of a method 300 of determining user interaction using steerable microphone arrays and a fixed-location microphone array. Each of a plurality of users has a steerable stereo microphone array, such as an ANC headset, that has a known orientation corresponding to the visual gazing direction of each such user. Each of the steerable arrays (in the ANC headsets) provides fixed broadside beamforming at 305, in which a beamformed output (or any type of spatially filtered output) is generated in the user look direction at 310 (i.e., in the direction the user of the steerable array is looking).

[0066] A fixed microphone array (such as in a smartphone) with an associated processor performs a direction of arrival (DOA) estimation at 320 in three dimensions (3D) around the fixed microphone array and separates the active speakers at 325. The number of active speakers is determined at 370, and a separate output for each active speaker (identified by an identification number for example) is generated at 380. In an implementation, speaker recognition and labeling of the active speakers may be performed at 330.

[0067] The similarity is measured between the separated speakers of the fixed array and the outputs of the steerable arrays at 340. Using the measured similarity and the

DOA estimation and the speaker IDs, a visualization of the user interaction (with speaker identity (ID) or participant ID) may be generated and displayed at 350. Each user's look direction may be provided to the fixed array as a smartphone coordinate for example, at 360.

[0068] A connection graph (also referred to as an interaction graph) may be generated which displays (a) who is talking and/or listening to whom and/or looking at whom, (b) who is dominating and/or leading the discussion of the group, and/or (c) who is bored, not participating, and/or quiet, for example. Real-time meeting analysis may be performed to assist the efficiency of the meeting and future meetings. Information such as time of meeting, place (e.g., meeting location), speaker identity or participant identity, meeting topic or subject matter, and number of participants, for example, may be displayed and used in the analysis.

[0069] Figure 4 is a diagram 400 of an example display 403 that may provide an indication of a user identity and which direction the user is looking. The user identity (participant ID 406) is displayed along with the direction that the user is looking (participant look direction 410). During a meeting, for example, this display of the participant look direction 410 may be generated and provided to an interested party, such as a meeting administrator or leader or supervisor, so that the interested party may see who the participant is looking at at various times during the meeting. Although only one participant ID 406 and participant look direction 410 is shown in the diagram 403, this is not intended to be limited. The interested party may receive such information for more than one participant, and such information may be displayed concurrently on one or more displays depending on the implementation. The data that is generated for display on the display 403 may be stored in a memory and retrieved and displayed at a later time, as well as being displayed in real-time.

[0070] Figure 5 is a diagram 415 of a user interface that may be generated and displayed on a display 418 and that indicates various user interactions and meeting data. Various types of information may be generated and displayed (e.g., in real-time during a meeting), such as the identifier (ID) of the participant who is talking 420, the ID of the participant(s) that is listening 422, and/or the ID of the participant(s) that is not participating 424 (e.g., not listening at the moment, not listening for more than a predetermined about of time or for at least a percentage of the meeting, looking somewhere other than the participant who is talking or looking in an another predetermined location or direction, etc.). During a meeting, for example, this display 4108 may be generated and provided to an interested party, such as a meeting administrator or leader or supervisor.

[0071] Additional data may be displayed on the display 418, such as the meeting time 426, the meeting location 428, the length of the meeting 430 (i.e., the duration), the meeting topic 432, and the number of meeting participants 434. Some or all of this data may be displayed. Additionally or alternatively, other data may be displayed, depending on the implementation, such as the IDs of all the participants and other statistics that may be generated as described further herein. The information and data that is generated for display on the display 418 may be stored in a memory and retrieved and displayed at a later time, as well as being displayed in real-time.

[0072] It is noted that a participant will be participating even if she is just listening at the meeting (and not speaking) because that participant's microphone (steerable microphone array) will still be picking up the sounds in the direction she is viewing while she is listening. Thus, even if a participant does not speak, there will still be sounds to analyze that are associated with her listening.

[0073] A user interface may be generated and displayed (e.g., on a smartphone display or other computing device display such as a display associated with a handset, a laptop, a tablet, a computer, a netbook, or a mobile computing device) that indicates the various user interactions during the meeting. Figure 4 is a diagram of an example display of a user interface 440 that may be generated and displayed (e.g., on a smartphone display 443) and that indicates various user interactions (e.g., during a meeting). In this example, the direction of each arrow 454 indicates who is looking at whom (only one arrow 454 is shown in this example, though a plurality of such arrows may be shown depending on the implementation and user interactions at a particular time). The thickness of each arrow indicates relatively how strong the interaction is (e.g., based on connected time, etc.). No arrow from or to a person indicates that the user is not involved in the group meeting. A percentage number may be displayed for a user which indicates a participation rate for the group meeting. An indicator 448 may be displayed to identify the leader of the meeting, and percentages 450, 452 may be determined and displayed to show how much of the discussion is directed to a person, and how much of the discussion is directed from the person, respectively. In an implementation, a color or highlighting may be used to indicate the leader of a group of participants.

[0074] In the example of Figure 6, John and Mark are interacting a lot, as indicated by the relatively big thick arrow 446. Mary is being quiet. Real-time meeting analysis (such as that described above with respect to Figures 4 and 5, and elsewhere herein) may be performed to assist the efficiency of the meeting. For example, because it looks like Mary is out of the conversation, John may encourage Mary to participate (e.g., by asking a question of Mary).

[0075] Social interaction plots may be accumulated over a time period (e.g., over a month, a year, etc.) to assess group dynamics or topic dynamics, for example. Figure 7 is a diagram 460 of an example display 462 that indicates various user interactions with respect to various topics 464. This information may be captured during one or more meetings, stored in a memory (or multiple memories), and displayed in one or more formats at a later time, e.g., during a historical analysis of data. Here, each participant ID 466 is listed along with their participation rates 468 for the various topics 464.

[0076] Thus, for example, Jane has a 20% participation rate in meetings about "Design", a 40% participation rate in meetings about "Code Walkthrough", and a 10% participation rate in meetings about "Documentation". This data may be used to determine which participants are most suited for, or interested in, a particular topic, for example, or which participants may need more encouragement with respect to a particular topic. Participation rates may be determined and based on one or more data items described herein, such as amount of time speaking at the meeting, amount of time paying attention at the meeting, amount of time listening at the meeting, etc. Although percentages are shown in Figure 7, any relative measuring, numbering, or indicating system or technique may be used to identify relative strengths and/or weaknesses in participating levels or rates.

[0077] An "L" in the diagram 460 is used as an example indicator to indicate which user participated most in a certain topic, thereby indicating a potential leader for that topic for example. Any indicator may be used, such as a color, highlighting, or a particular symbol. In this example, John is the most participating in Design, Jane is the most participating in Code Walkthrough, and Mary is the most participating in Documentation. Accordingly, they may be identified as potential leaders in the respective topics.

[0078] Additionally, a personal time line with an interaction history may be generated for one or more meeting participants. Thus, not only a single snapshot or period of time during a meeting may be captured, analyzed, and information pertaining to it displayed (either in real-time or later offline), but also history over time may be stored (e.g., in a memory of a computing device such as a smartphone or any type of computing device, such as a handset, a laptop, a tablet, a computer, a netbook, or a mobile computing device), analyzed, and displayed (e.g., in a calendar or other display of a computing device such as a smartphone any type of computing device, such as a handset, a laptop, a tablet, a computer, a netbook, or a mobile computing device).

[0079] Figure 8 is a diagram 470 of an example display 472 that indicates various user interactions over time, that may be used for historical analysis, e.g., after one or more meetings. Here, a user identifier 474 is provided, along with information such as the meeting date and the meeting topic. The information 478 on this display 472 is provided over time 476. It shows information 478, for each period or instant of time, such as who the user was looking at that period or instant of time, whether the user was speaking then, and the percentage of meeting participants that were looking at the user at the period or instant of time. This information 478 can be determined at predetermined times during a meeting (e.g., every minute, every 5 minutes, etc.), or determined as an average or other weighted determination over particular periods of time, for example. This information is provided as an example only and is not meant to be limiting; additional or alternative information can be generated and displayed as information 478.

[0080] The information displayed in Figure 8 can be used for meeting analysis and user analysis. Thus, in Figure 8, it may be determined that the user Jane typically looks at Mary or Mark when Jane is not speaking, but Jane looks at John when Jane is speaking. Figure 8 also indicates that when Jane is not speaking, the percentage of participants looking at Jane is zero, but this percentage increases as Jane is speaking.

[0081] Interaction statistics may also be generated, stored, analyzed, and displayed. For example, the evolution of interaction between people can be tracked and displayed. Recursive weighting over time may be used (e.g., 0.9*historical data + 0.1*current data), such that as data gets older, it becomes less relevant, with the most current data being weighted the highest (or vice versa). In this manner, a user may be able to see which people he or others are networking with more than others. Additional statistics may be factored into the analysis to provide more accurate interaction information. For example, interaction information obtained from email exchanges or other communication may be used (combined with) the meeting, history, and/or participant interaction data to provide additional (e.g., more accurate) interaction information.

[0082] Figure 9 is a diagram 480 of another example display 482 that indicates various user interactions over time. Here, a user Jane is identified along with an interaction scale 488 and a time period. The diagram 480 shows other user IDs 484 and a listing of months 486 in the past. The interaction scale in this example ranges from 0 to 10, with 0 representing no interaction and 10 representing a very strong interaction between the identified user and Jane in each of the months 486. This information may be generated and provided as historical data and used, e.g., by a meeting participant or a leader or supervisor to view and analyze the various user interactions over time, e.g., to see who is most strongly interacting with whom when.

[0083] As another example, online learning monitoring may be performed to determine whether a student in a remote site is actively participating or not. Likewise, an application for video games with participant interaction is also contemplated in which there may be immediate recognition of where the users are looking among the possible sound event locations.

[0084] Figure 10 is an operational flow of an implementation of a method 500, and uses cross-correlation as an exemplary measure although any similarity measurement technique may be used. At 503, the fixed microphone array provides a number of active speakers N and the active speakers' separated speech signals. One signal (the sound) is received by the fixed microphone array. The output of the fixed microphone array comprises beams, one beam corresponding to each participant. Thus, a separate output is associated with each participant. At 510, the steerable microphone array provides the user's look direction. For each user, the individual user's output is correlated with each of the beamforms (or other spatially filtered output) that are outputted from the fixed microphone array.

[0085] Location mapping may be generated using this information, at 515. Information pertaining to when a user turns to someone and looks at them may be leveraged. A well known classic correlation equation, such as that shown at 506, may be used as shown, where E is equal to the expectation value and c is the correlation value. Whenever there is a maximum peak, that is the angle of strong correlation. In an implementation, the maximum allowable time shift may be predetermined using a physical constraint or system complexity. For example, the time delay between steerable microphones and fixed microphones can be measured and used, when only the user, who wears the steerable array, is active. Note that the conventional frame length 20 ms corresponds to almost 7 meters. The angle θ is the relative angle at which the active speaker is located relative to the listening user. The angle θ may be determined between the fixed array and the steerable array, at 513.

[0086] Figure 11 is an operational flow of an implementation of a method 520 of measuring similarity, and uses cross-cumulant as an exemplary measure although any similarity measurement technique may be used. The fixed microphone array provides a number of active speakers N and the active speakers' separated speech signals, at 523. One signal (the sound) is received by the fixed microphone array. The output of the fixed microphone array comprises beams, one beam corresponding to each participant. Thus, a separate output is associated with each participant. The steerable microphone array provides the user's look direction, at 530. For each user, the individual user's output is correlated with each of the beamforms (or other spatially filtered output) that is outputted from the fixed microphone array.

[0087] Location mapping may be generated using this information, at 525. Information pertaining to when a user turns to someone and looks at them may be leveraged. A well known classic cumulant equation, shown at 526, may be used as shown, where E is equal to the expectation value and c is the correlation value. Whenever there is a maximum peak, that is the angle of strong correlation. The angle θ is the relative angle at which the active speaker is located relative to the listening user. The angle θ may be determined between the fixed array and the steerable array, at 513.

[0088] It is noted that any similarity or correlation technique may be used. Regarding a possible similarity measure, virtually any distance metric(s) may be used such as, but not limited to the well known techniques of: (1) least square fit with allowable time adjustment: time-domain or frequency-domain; (2) feature based approach: using linear prediction coding (LPC) or mel-frequency cepstral coefficients (MFCC); and (3) higher order based approach: cross-cumulant, empirical Kullback-Leibler Divergence, or Itakura–Saito distance.

[0089] Figure 12 is an operational flow of an implementation of a method 540 of measuring similarity using time-domain least squares fit, and Figure 13 is an operational flow of an implementation of a method 550 of measuring similarity using frequency-domain least squares fit. The method 540, using a time-domain least squares fit, is similar to the method 520 of Figure 11 described above, except that instead of using a cumulant equation of 526, a time domain equation shown at 542 may be used as shown. Similarly, the method 550 is similar to the method 520 of Figure 11 but instead of using energy normalization, uses a fast Fourier transform (FFT) in conjunction with the frequency domain equation shown at 552.

[0090] Figure 14 is an operational flow of an implementation of a method 560 of measuring similarity using Itakura-Saito distance. This technique is similar to the FFT technique of Figure 13, but uses the equation shown at 562. Figure 15 is an operational flow of an implementation of a method 570 of measuring similarity using a feature based approach. Feature extraction is performed, as shown at 573 and 575, and used in conjunction with the other operations 503, 510, 513, and 515 of Figure 10, and the equation shown at 572.

[0091] In an implementation, the correlation or similarity between the audio beams of the separated speakers of the fixed microphone array and the outputs of the steerable microphone arrays may be used to zoom into a targeted speaker. This type of collaborative zooming may provide a user interface for zooming into a desired speaker.

[0092] In other words, collaborative zooming may be performed wherein a user interface is provided for multiple users with multiple devices for zooming into a target speaker by just looking at the target speaker. Beamforming may be produced at the targeted person via either the headsets or handsets such that all available resources of multiple devices can be combined for collaborative zooming, thereby enhancing the look direction of the targeted person.

[0093] For example, a user may look at a target person, and beamforming may be produced at the targeted person by either using the headset or a handset (whichever is closer to the target person). This may be achieved by using a device that includes a hidden camera with two microphones. When multiple users of multiple devices look at the target person, the camera(s) can visually focus on the person. In addition, the device(s) can audibly focus (i.e., zoom in on) the person by using (e.g., all) available microphones to enhance the look direction of the target person.

[0094] Additionally, the target person can be audibly zoomed in on by nulling out other speakers and enhancing the target person's voice. The enhancement can also be done using a headset or handset, whichever is closer to the target person.

[0095] An exemplary user interface display 600 is shown in Figure 16. The display (e.g., displayed on a smartphone display 610 or other display device) shows the active user location 620 and an associated energy 630. Figure 17 shows an exemplary user interface display to show collaborative zooming on the display, in which Speaker 1 is zoomed in on as shown in the display 660 from the initial display 650.

[0096] Figure 18 is an operational flow of an implementation of a method 700 for zooming into a target person. As in Figure 3, a steerable array 705 (in an ANC headset) provides fixed broadside beamforming at 710, in which a beamformed output is generated in the user look direction (i.e., in the direction the user of the steerable array is looking). A fixed microphone array 707 (such as in a smartphone) with an associated processor performs a DOA estimation in three dimensions around the fixed microphone array and separates the active speakers, at 720. The number of active speakers is determined, and a separate output for each active speaker (identified by an identification number for example) is generated.

[0097] In an implementation, speaker recognition and labeling of the active speakers may be performed at 730. At 750, a correlation or similarity is determined between the separated speakers of the fixed array and the outputs of the steerable arrays. Using the correlation or similarity measurement and the speakers' IDs, a target user can be detected, localized, and zoomed into, at 760.

[0098] The user can be replaced with a device, such as a hidden camera with two microphones, and just by looking at the targeted person, the targeted person can be focused on with zooming by audition as well as by vision.

[0099] A camcorder application with multiple devices is contemplated. The look direction is known, and all available microphones of other devices may be used to enhance the look direction source.

[00100] In an implementation, the correlation or similarity between the audio beams of the separated speakers of the fixed array and the outputs of steerable arrays may be used to adaptively form a better beam for a targeted speaker. In this manner, the fixed microphones beamformer may be adaptively refined, such that new look directions can be adaptively generated by a fixed beamformer.

[00101] For example, the headset microphone array's beamformer output can be used as a reference to refine the look direction of fixed microphone array's beamformer. The correlation or similarity between the headset beamformer output and the current fixed microphone array beamformer output may be compared with the correlation or similarity between the headset beamformer output and the fixed microphone array beamformer outputs with slightly moved look directions.

[00102] Figure 19 shows an example user interface display 800 with additional candidate look directions 810. By leveraging the correlation or similarity between the headset beamformer output with the original fixed microphone beamformer outputs 820, as shown in Figure 19, new candidate look directions by a fixed beamformer can be generated. Using this technique, the headset microphone beamformer output can be used as a reference to refine the look direction of the fixed microphone beamformer. For example, speaker 1 in Figure 19 may be speaking, and as he speaks new candidate look directions can be adaptively formed.

[00103] Figure 20 is an operational flow of an implementation of a method 900 for adaptively refining beams for a targeted speaker. As in Figure 3, a steerable array 905 (for example, in an ANC headset) provides fixed broadside beamforming at 910, in which a beamformed output is generated in the user look direction (i.e., in the direction the user of the steerable array is looking). A fixed microphone array 907 (such as in a smartphone) with an associated processor performs a DOA estimation in three

dimensions around the fixed microphone array and separates the active speakers, at 920. The number of active speakers is determined, and a separate output for each active speaker (identified by an identification number for example) is generated. As with Figure 18, a correlation or similarity is determined between the separated speakers of the fixed array and the outputs of the steerable arrays, at 950.

[00104] Continuing with Figure 20, the determined correlation or similarity is used to increase the angular resolution near the DOAs of the active users, and a separation of the active speakers is again performed, at 960. Using the increased angular resolution and the outputs of the steerable arrays, another correlation or similarity measure is determined between the separated speakers of the fixed array and the outputs of the steerable arrays, at 970. This correlation or similarity measure may then be used to zoom into a target speaker, at 980.

[00105] It is a challenge to provide a method for estimating a three-dimensional direction of arrival (DOA) for each frame of an audio signal for concurrent multiple sound events that is sufficiently robust under background noise and reverberation. Robustness can be obtained by maximizing the number of reliable frequency bins. It may be desirable for such a method to be suitable for arbitrarily shaped microphone array geometry, such that specific constraints on microphone geometry may be avoided. A pair-wise 1-D approach as described herein can be appropriately incorporated into any geometry.

[00106] A solution may be implemented for such a generic speakerphone application or far-field application. Such an approach may be implemented to operate without a microphone placement constraint. Such an approach may also be implemented to track sources using available frequency bins up to Nyquist frequency and down to a lower frequency (e.g., by supporting use of a microphone pair having a larger inter-microphone distance). Rather than being limited to a single pair for tracking, such an approach may be implemented to select a best pair among all available pairs. Such an approach may be used to support source tracking even in a far-field scenario, up to a distance of three to five meters or more, and to provide a much higher DOA resolution. Other potential features include obtaining an exact 2-D representation of an active source. For best results, it may be desirable that each source is a sparse broadband audio source, and that each frequency bin is mostly dominated by no more than one source.

[00107] For a signal received by a pair of microphones directly from a point source in a particular DOA, the phase delay differs for each frequency component and also depends on the spacing between the microphones. The observed value of the phase delay at a particular frequency bin may be calculated as the inverse tangent of the ratio of the imaginary term of the complex FFT coefficient to the real term of the complex FFT coefficient. As shown in Figure 21, the phase delay value $\Delta\varphi_f$ at a particular frequency f may be related to source DOA under a far-field (i.e., plane-wave) assumption as $\Delta\varphi_f = 2\pi f \frac{d \sin\theta}{c}$, where d denotes the distance between the microphones (in m), θ denotes the angle of arrival (in radians) relative to a direction that is orthogonal to the array axis, f denotes frequency (in Hz), and c denotes the speed of sound (in m/s). For the ideal case of a single point source with no reverberation, the ratio of phase delay to frequency $\Delta\varphi/f$ will have the same value $2\pi \frac{d \sin\theta}{c}$ over all frequencies.

[00108] Such an approach is limited in practice by the spatial aliasing frequency for the microphone pair, which may be defined as the frequency at which the wavelength of the signal is twice the distance d between the microphones. Spatial aliasing causes phase wrapping, which puts an upper limit on the range of frequencies that may be used to provide reliable phase delay measurements for a particular microphone pair. Figure 23 shows plots of unwrapped phase delay vs. frequency for four different DOAs, and Figure 24 shows plots of wrapped phase delay vs. frequency for the same DOAs, where the initial portion of each plot (i.e., until the first wrapping occurs) are shown in bold. Attempts to extend the useful frequency range of phase delay measurement by unwrapping the measured phase are typically unreliable.

[00109] Instead of phase unwrapping, a proposed approach compares the phase delay as measured (e.g., wrapped) with pre-calculated values of wrapped phase delay for each of an inventory of DOA candidates. Figure 25 shows such an example that includes angle-vs.-frequency plots of the (noisy) measured phase delay values (gray) and the phase delay values for two DOA candidates of the inventory (solid and dashed lines), where phase is wrapped to the range of pi to minus pi. The DOA candidate that is best matched to the signal as observed may then be determined by calculating, for each DOA candidate θ_i, a corresponding error e_i between the phase delay values $\Delta\varphi_{i\_f}$ for the i-th DOA candidate and the observed phase delay values $\Delta\varphi_{ob\_f}$ over a range of frequency components f, and identifying the DOA candidate value that corresponds to the minimum error. In one example, the error e_i is expressed as $\left\|\Delta\varphi_{ob\_f} - \Delta\varphi_{i\_f}\right\|_f^2$, i.e. as the sum

$$e_i = \sum_{f \in F} \left(\Delta\varphi_{ob\_f} - \Delta\varphi_{i\_f}\right)^2$$

of the squared differences between the observed and candidate phase delay values over a desired range or other set F of frequency components. The phase delay values $\Delta\varphi_{i\_f}$ for each DOA candidate θ_i may be calculated before run-time (e.g., during design or manufacture), according to known values of c and d and the desired range of frequency components f, and retrieved from storage during use of the device. Such a pre-calculated inventory may be configured to support a desired angular range and resolution (e.g., a uniform resolution, such as one, two, five, or ten degrees; or a desired nonuniform resolution) and a desired frequency range and resolution (which may also be uniform or nonuniform).

[00110] It may be desirable to calculate the error e_i across as many frequency bins as possible to increase robustness against noise. For example, it may be desirable for the error calculation to include terms from frequency bins that are beyond the spatial aliasing frequency. In a practical application, the maximum frequency bin may be limited by other factors, which may include available memory, computational complexity, strong reflection by a rigid body at high frequencies, etc.

[00111] A speech signal is typically sparse in the time-frequency domain. If the sources are disjoint in the frequency domain, then two sources can be tracked at the same time. If the sources are disjoint in the time domain, then two sources can be tracked at the same frequency. It may be desirable for the array to include a number of microphones that is at least equal to the number of different source directions to be distinguished at any one time. The microphones may be omnidirectional (e.g., as may be typical for a cellular telephone or a dedicated conferencing device) or directional (e.g., as may be typical for a device such as a set-top box).

[00112] Such multichannel processing is generally applicable, for example, to source tracking for speakerphone applications. Such a technique may be used to calculate a DOA estimate for a frame of the received multichannel signal. Such an approach may calculate, at each frequency bin, the error for each candidate angle with respect to the observed angle, which is indicated by the phase delay. The target angle at that frequency bin is the candidate having the minimum error. In one example, the error is then summed across the frequency bins to obtain a measure of likelihood for the candidate. In another example, one or more of the most frequently occurring target DOA candidates across all frequency bins is identified as the DOA estimate (or estimates) for a given frame.

[00113] Such a method may be applied to obtain instantaneous tracking results (e.g., with a delay of less than one frame). The delay is dependent on the FFT size and the degree of overlap. For example, for a 512-point FFT with a 50% overlap and a sampling frequency of 16 kHz, the resulting 256-sample delay corresponds to sixteen milliseconds. Such a method may be used to support differentiation of source directions typically up to a source-array distance of two to three meters, or even up to five meters.

[00114] The error may also be considered as a variance (i.e., the degree to which the individual errors deviate from an expected value). Conversion of the time-domain received signal into the frequency domain (e.g., by applying an FFT) has the effect of averaging the spectrum in each bin. This averaging is even more obvious if a subband representation is used (e.g., mel scale or Bark scale). Additionally, it may be desirable to perform time-domain smoothing on the DOA estimates (e.g., by applying as recursive smoother, such as a first-order infinite-impulse-response filter).

[00115] It may be desirable to reduce the computational complexity of the error calculation operation (e.g., by using a search strategy, such as a binary tree, and/or applying known information, such as DOA candidate selections from one or more previous frames).

[00116] Even though the directional information may be measured in terms of phase delay, it is typically desired to obtain a result that indicates source DOA. Consequently, it may be desirable to calculate the error in terms of DOA rather than in terms of phase delay.

[00117] An expression of error e_i in terms of DOA may be derived by assuming that an expression for the observed wrapped phase delay as a function of DOA, such as

$$\Psi_{f\_wr}(\theta) = \mathrm{mod}\left(-2\pi f \frac{d \sin\theta}{c} + \pi, 2\pi\right) - \pi,$$

is equivalent to a corresponding expression for unwrapped phase delay as a function of DOA, such as

$$\Psi_{f\_un}(\theta) = -2\pi f \frac{d \sin\theta}{c},$$

except near discontinuities that are due to phase wrapping. The error e_i may then be expressed as

$$e_i = \left\|\Psi_{f\_wr}(\theta_{ob}) - \Psi_{f\_wr}(\theta_i)\right\|_f^2 \equiv \left\|\Psi_{f\_un}(\theta_{ob}) - \Psi_{f\_un}(\theta_i)\right\|_f^2,$$

where the difference between the observed and candidate phase delay at frequency f is expressed in terms of DOA as

$$\Psi_{f\_un}(\theta_{ob}) - \Psi_{f\_un}(\theta_i) = \frac{-2\pi f d}{c}\left(\sin\theta_{ob\_f} - \sin\theta_i\right).$$

[00118] Perform a Taylor series expansion to obtain the following first-order approximation:

$$\frac{-2\pi f d}{c}\left(\sin\theta_{ob\_f} - \sin\theta_i\right) \approx \left(\theta_{ob\_f} - \theta_i\right)\frac{-2\pi f d}{c}\cos\theta_i,$$

which is used to obtain an expression of the difference between the DOA θ_{ob_f} as observed at frequency f and DOA candidate θ_i:

$$\left(\theta_{ob\_f} - \theta_i\right) \cong \frac{\Psi_{f\_un}(\theta_{ob}) - \Psi_{f\_un}(\theta_i)}{\frac{2\pi f d}{c}\cos\theta_i}.$$

This expression may be used, with the assumed equivalence of observed wrapped phase delay to unwrapped phase delay, to express error e_i in terms of DOA:

$$e_i = \|\theta_{ob} - \theta_i\|_f^2 \cong \frac{\left\|\Psi_{f\_wr}(\theta_{ob}) - \Psi_{f\_wr}(\theta_i)\right\|_f^2}{\left\|\frac{2\pi f d}{c}\cos\theta_i\right\|_f^2},$$

where the values of $[\Psi_{f\_wr}(\theta_{ob}), \Psi_{f\_wr}(\theta_i)]$ are defined as $[\Delta\varphi_{ob\_f}, \Delta\varphi_{i\_f}]$.

[00119] To avoid division with zero at the endfire directions (θ = +/– 90°), it may be desirable to perform such an expansion using a second-order approximation instead, as in the following:

$$|\theta_{ob} - \theta_i| \cong \begin{cases} |-C/B|, & \theta_i = 0 \text{ (broadside)} \\ \left|\frac{-B + \sqrt{B^2 - 4AC}}{2A}\right|, & \text{otherwise} \end{cases},$$

where $A = (\pi f d \sin\theta_i)/c$, $B = (-2\pi f d \cos\theta_i)/c$, and

$$C = -\left(\Psi_{f\_un}(\theta_{ob}) - \Psi_{f\_un}(\theta_i)\right).$$

As in the first-order example above, this expression may be used, with the assumed equivalence of observed wrapped phase delay to unwrapped phase delay, to express error e_i in terms of DOA as a function of the observed and candidate wrapped phase delay values.

[00120] As shown in Figure 27, a difference between observed and candidate DOA for a given frame of the received signal may be calculated in such manner at each of a plurality of frequencies f of the received microphone signals (e.g., $\forall f \in F$) and for each of a plurality of DOA candidates θ_i. As demonstrated in Figure 28, a DOA estimate for a given frame may be determined by summing the squared differences for each candidate across all frequency bins in the frame to obtain the error e_i and selecting the DOA candidate having the minimum error. Alternatively, as demonstrated in Figure 29, such differences may be used to identify the best-matched (i.e. minimum squared difference) DOA candidate at each frequency. A DOA estimate for the frame may then be determined as the most frequent DOA across all frequency bins.

[00121] As shown in Figure 31, an error term may be calculated for each candidate angle i and each of a set F of frequencies for each frame k. It may be desirable to indicate a likelihood of source activity in terms of a calculated DOA difference or error. One example of such a likelihood L may be expressed, for a particular frame, frequency, and angle, as

$$L(i,f,k) = \frac{1}{|\theta_{ob} - \theta_i|_{f,k}^2}. \quad (1)$$

[00122] For expression (1), an extremely good match at a particular frequency may cause a corresponding likelihood to dominate all others. To reduce this susceptibility, it may be desirable to include a regularization term λ, as in the following expression:

$$L(i,f,k) = \frac{1}{|\theta_{ob} - \theta_i|_{f,k}^2 + \lambda}. \quad (2)$$

[00123] Speech tends to be sparse in both time and frequency, such that a sum over a set of frequencies F may include results from bins that are dominated by noise. It may be desirable to include a bias term β, as in the following expression:

$$L(i,f,k) = \frac{1}{|\theta_{ob} - \theta_i|_{f,k}^2 + \lambda} - \beta. \quad (3)$$

The bias term, which may vary over frequency and/or time, may be based on an assumed distribution of the noise (e.g., Gaussian). Additionally or alternatively, the bias term may be based on an initial estimate of the noise (e.g., from a noise-only initial frame). Additionally or alternatively, the bias term may be updated dynamically based on information from noise-only frames, as indicated, for example, by a voice activity detection module.

[00124] The frequency-specific likelihood results may be projected onto a (frame, angle) plane to obtain a DOA estimation per frame $\theta_{est\_k} = \max_i \sum_{f \in F} L(i,f,k)$ that is robust to noise and reverberation because only target dominant frequency bins contribute to the estimate. In this summation, terms in which the error is large have values that approach zero and thus become less significant to the estimate. If a directional source is dominant in some frequency bins, the error value at those frequency bins will be nearer to zero for that angle. Also, if another directional source is dominant in other frequency bins, the error value at the other frequency bins will be nearer to zero for the other angle.

[00125] The likelihood results may also be projected onto a (frame, frequency) plane to indicate likelihood information per frequency bin, based on directional membership (e.g., for voice activity detection). This likelihood may be used to indicate likelihood of speech activity. Additionally or alternatively, such information may be used, for example, to support time- and/or frequency-selective masking of the received signal by classifying frames and/or frequency components according to their direction of arrival.

[00126] An anglogram representation is similar to a spectrogram representation. An anglogram may be obtained by plotting, at each frame, a likelihood of the current DOA candidate at each frequency.

[00127] A microphone pair having a large spacing is typically not suitable for high frequencies, because spatial aliasing begins at a low frequency for such a pair. A DOA estimation approach as described herein, however, allows the use of phase delay measurements beyond the frequency at which phase wrapping begins, and even up to the Nyquist frequency (i.e., half of the sampling rate). By relaxing the spatial aliasing constraint, such an approach enables the use of microphone pairs having larger inter-microphone spacings. As an array with a large inter-microphone distance typically provides better directivity at low frequencies than an array with a small inter-microphone distance, use of a larger array typically extends the range of useful phase delay measurements into lower frequencies as well.

[00128] The DOA estimation principles described herein may be extended to multiple microphone pairs in a linear array (e.g., as shown in Figure 22). One example of such an application for a far-field scenario is a linear array of microphones arranged along the margin of a television or other large-format video display screen (e.g., as shown in Figure 26). It may be desirable to configure such an array to have a nonuniform (e.g., logarithmic) spacing between microphones, as in the examples of Figures 22 and 26.

[00129] For a far-field source, the multiple microphone pairs of a linear array will have essentially the same DOA. Accordingly, one option is to estimate the DOA as an average of the DOA estimates from two or more pairs in the array. However, an averaging scheme may be affected by mismatch of even a single one of the pairs, which may reduce DOA estimation accuracy. Alternatively, it may be desirable to select, from among two or more pairs of microphones of the array, the best microphone pair for each frequency (e.g., the pair that gives the minimum error e_i at that frequency), such that different microphone pairs may be selected for different frequency bands. At the spatial aliasing frequency of a microphone pair, the error will be large. Consequently, such an approach will tend to automatically avoid a microphone pair when the frequency is close to its wrapping frequency, thus avoiding the related uncertainty in the DOA estimate. For higher-frequency bins, a pair having a shorter distance between the microphones will typically provide a better estimate and may be automatically favored, while for lower-frequency bins, a pair having a larger distance between the microphones will typically provide a better estimate and may be automatically favored. In the four-microphone example shown in Figure 22, six different pairs of microphones are possible (i.e., $\binom{4}{2} = 6$).

[00130] In one example, the best pair for each axis is selected by calculating, for each frequency f, PxI values, where P is the number of pairs, I is the size of the inventory, and each value e_{pi} is the squared absolute difference between the observed angle θ_{pf} (for pair p and frequency f) and the candidate angle θ_{if}. For each frequency f, the pair p that corresponds to the lowest error value e_{pi} is selected. This error value also indicates the best DOA candidate θ_i at frequency f (as shown in Figure 30).

[00131] The signals received by a microphone pair may be processed as described herein to provide an estimated DOA, over a range of up to 180 degrees, with respect to the axis of the microphone pair. The desired angular span and resolution may be arbitrary within that range (e.g. uniform (linear) or nonuniform (nonlinear), limited to selected sectors of interest, etc.). Additionally or alternatively, the desired frequency span and resolution may be arbitrary (e.g. linear, logarithmic, mel-scale, Bark-scale, etc.).

[00132] In the model shown in Figure 22, each DOA estimate between 0 and +/– 90 degrees from a microphone pair indicates an angle relative to a plane that is orthogonal to the axis of the pair. Such an estimate describes a cone around the axis of the pair, and the actual direction of the source along the surface of this cone is indeterminate. For example, a DOA estimate from a single microphone pair does not indicate whether the source is in front of or behind the microphone pair. Therefore, while more than two microphones may be used in a linear array to improve DOA estimation performance across a range of frequencies, the range of DOA estimation supported by a linear array is typically limited to 180 degrees.

[00133] The DOA estimation principles described herein may also be extended to a two-dimensional (2-D) array of microphones. For example, a 2-D array may be used to extend the range of source DOA estimation up to a full 360^0 (e.g., providing a similar range as in applications such as radar and biomedical scanning). Such an array may be used in a speakerphone application, for example, to support good performance even for arbitrary placement of the telephone relative to one or more sources.

[00134] The multiple microphone pairs of a 2-D array typically will not share the same DOA, even for a far-field point source. For example, source height relative to the plane of the array (e.g., in the z-axis) may play an important role in 2-D tracking. Figure 32 shows an example of a speakerphone application in which the x-y plane as defined by the microphone axes is parallel to a surface (e.g., a tabletop) on which the telephone is placed. In this example, the source is a person speaking from a location that is along the x axis but is offset in the direction of the z axis (e.g., the speaker's mouth is above the tabletop). With respect to the x-y plane as defined by the microphone array, the direction of the source is along the x axis, as shown in Figure 32. The microphone pair along the y axis estimates a DOA of the source as zero degrees from the x-z plane. Due to the height of the speaker above the x-y plane, however, the microphone pair along the x axis estimates a DOA of the source as 30^0 from the x axis (i.e., 60 degrees from the y-z plane), rather than along the x axis. Figures 34 and 35 shows two views of the cone of confusion associated with this DOA estimate, which causes an ambiguity in the estimated speaker direction with respect to the microphone axis.

[00135] An expression such as

$$\left[\tan^{-1}\left(\frac{\sin\theta_1}{\sin\theta_2}\right), \tan^{-1}\left(\frac{\sin\theta_2}{\sin\theta_1}\right)\right], \qquad (4)$$

where θ_1 and θ_2 are the estimated DOA for pair 1 and 2, respectively, may be used to project all pairs of DOAs to a 360^0 range in the plane in which the three microphones are located. Such projection may be used to enable tracking directions of active speakers over a 360^0 range around the microphone array, regardless of height difference. Applying the expression above to project the DOA estimates (0^0, 60^0) of Figure 32 into the x-y plane produces

$$\left[\tan^{-1}\left(\frac{\sin 0^\circ}{\sin 60^\circ}\right), \tan^{-1}\left(\frac{\sin 60^\circ}{\sin 0^\circ}\right)\right] = (0^\circ, 90^\circ),$$

which may be mapped to a combined directional estimate (e.g., an azimuth) of 270^0 as shown in Figure 33.

[00136] In a typical use case, the source will be located in a direction that is not projected onto a microphone axis. Figures 37-40 show such an example in which the source is located above the plane of the microphones. In this example, the DOA of the source signal passes through the point (x,y,z) = (5,2,5). Figure 37 shows the x-y plane as viewed from the +z direction, Figures 38 and 40 show the x-z plane as viewed from the direction of microphone MC30, and Figure 39 shows the y-z plane as viewed from the direction of microphone MC10. The shaded area in Figure 37 indicates the cone of confusion CY associated with the DOA θ_1 as observed by the y-axis microphone pair MC20-MC30, and the shaded area in Figure 38 indicates the cone of confusion CX associated with the DOA θ_2 as observed by the x-axis microphone pair MC10-MC20. In Figure 39, the shaded area indicates cone CY, and the dashed circle indicates the intersection of cone CX with a plane that passes through the source and is orthogonal to the x axis. The two dots on this circle that indicate its intersection with cone CY are the candidate locations of the source. Likewise, in Figure 40 the shaded area indicates cone CX, the dashed circle indicates the intersection of cone CY with a plane that passes through the source and is orthogonal to the y axis, and the two dots on this circle that indicate its intersection with cone CX are the candidate locations of the source. It may be seen that in this 2-D case, an ambiguity remains with respect to whether the source is above or below the x-y plane.

[00137] For the example shown in Figures 37-40, the DOA observed by the x-axis microphone pair MC10-MC20 is $\theta_2 = \tan^{-1}\left(-5/\sqrt{25+4}\right) \approx -42.9°$, and the DOA observed by the y-axis microphone pair MC20-MC30 is $\theta_1 = \tan^{-1}\left(-2/\sqrt{25+25}\right) \approx -15.8°$. Using expression (4) to project these directions into the x-y plane produces the magnitudes (21.8°, 68.2°) of the desired angles relative to the x and y axes, respectively, which corresponds to the given source location (x,y,z) = (5,2,5). The signs of the observed angles indicate the x-y quadrant in which the source is located, as shown in Figure 36.

[00138] In fact, almost 3D information is given by a 2D microphone array, except for the up-down confusion. For example, the directions of arrival observed by microphone pairs MC10-MC20 and MC20-MC30 may also be used to estimate the magnitude of the angle of elevation of the source relative to the x-y plane. If d denotes the vector from microphone MC20 to the source, then the lengths of the projections of vector d onto the x-axis, the y-axis, and the x-y plane may be expressed as $d\sin(\theta_2)$, $d\sin(\theta_1)$, and $d\sqrt{\sin^2(\theta_1)+\sin^2(\theta_2)}$, respectively. The magnitude of the angle of elevation may then be estimated as $\hat{\theta}_h = \cos^{-1}\sqrt{\sin^2(\theta_1)+\sin^2(\theta_2)}$.

[00139] Although the microphone pairs in the particular examples of Figures 32-33 and 37-40 have orthogonal axes, it is noted that for microphone pairs having non-orthogonal axes, expression (4) may be used to project the DOA estimates to those non-orthogonal axes, and from that point it is straightforward to obtain a representation of the combined directional estimate with respect to orthogonal axes. Figure 41 shows a example of microphone array MC10-MC20-MC30 in which the axis 1 of pair MC20-MC30 lies in the x-y plane and is skewed relative to the y axis by a skew angle θ_0.

[00140] Figure 42 shows an example of obtaining a combined directional estimate in the x-y plane with respect to orthogonal axes x and y with observations (θ_1, θ_2) from an array as shown in Figure 41. If d denotes the vector from microphone MC20 to the source, then the lengths of the projections of vector d onto the x-axis and axis 1 may be expressed as $d\sin(\theta_2)$ and $d\sin(\theta_1)$, respectively. The vector (x,y) denotes the projection of vector d onto the x-y plane. The estimated value of x is known, and it remains to estimate the value of y.

[00141] The estimation of y may be performed using the projection $p_1 = (d\sin\theta_1\sin\theta_0, d\sin\theta_1\cos\theta_0)$ of vector (x,y) onto axis 1. Observing that the difference between vector (x,y) and vector p_1 is orthogonal to p_1, calculate y as

$$y = d\frac{\sin\theta_1 - \sin\theta_2\sin\theta_0}{\cos\theta_0}.$$

The desired angles of arrival in the x-y plane, relative to the orthogonal x and y axes, may then be expressed respectively as

$$\left(\tan^{-1}\left(\frac{y}{x}\right), \tan^{-1}\left(\frac{x}{y}\right)\right)$$
$$= \left(\tan^{-1}\left(\frac{\sin\theta_1 - \sin\theta_2\sin\theta_0}{\sin\theta_2\cos\theta_0}\right), \tan^{-1}\left(\frac{\sin\theta_2\cos\theta_0}{\sin\theta_1 - \sin\theta_2\sin\theta_0}\right)\right).$$

[00142] Extension of DOA estimation to a 2-D array is typically well-suited to and sufficient for a speakerphone application. However, further extension to an N-dimensional array is also possible and may be performed in a straightforward manner. For tracking applications in which one target is dominant, it may be desirable to select N pairs for representing N dimensions. Once a 2-D result is obtained with a particular microphone pair, another available pair can be utilized to increase degrees of freedom. For example, Figures 37-42 illustrate use of observed DOA estimates from different microphone pairs in the x-y plane to obtain an estimate of the source direction as projected into the x-y plane. In the same manner, observed DOA estimates from an x-axis microphone pair and a z-axis microphone pair (or other pairs in the x-z plane) may be used to obtain an estimate of the source direction as projected into the x-z plane, and likewise for the y-z plane or any other plane that intersects three or more of the microphones.

[00143] Estimates of DOA error from different dimensions may be used to obtain a combined likelihood estimate, for example, using an expression such as

$$\frac{1}{\max\left(\left|\theta-\theta_{0,1}\right|_{f,1}^2, \left|\theta-\theta_{0,2}\right|_{f,2}^2\right)+\lambda} \text{ or } \frac{1}{\text{mean}\left(\left|\theta-\theta_{0,1}\right|_{f,1}^2, \left|\theta-\theta_{0,2}\right|_{f,2}^2\right)+\lambda},$$

where $\theta_{0,i}$ denotes the DOA candidate selected for pair i. Use of the maximum among the different errors may be desirable to promote selection of an estimate that is close to the cones of confusion of both observations, in preference to an estimate that is close to only one of the cones of confusion and may thus indicate a false peak. Such a combined result may be used to obtain a (frame, angle) plane, as described herein, and/or a (frame, frequency) plot, as described herein.

[00144] The DOA estimation principles described herein may be used to support selection among multiple speakers. For example, location of multiple sources may be combined with a manual selection of a particular speaker (e.g., push a particular button to select a particular corresponding user) or automatic selection of a particular speaker (e.g., by speaker recognition). In one such application, a telephone is configured to recognize the voice of its owner and to automatically select a direction corresponding to that voice in preference to the directions of other sources.

[00145] A source DOA may be easily defined in 1-D, e.g. from -90^0 to $+90^0$. For more than two microphones at arbitrary relative locations, it is proposed to use a straightforward extension of 1-D as described above, e.g. (θ_1, θ_2) in two-pair case in 2-D, $(\theta_1, \theta_2, \theta_3)$ in three-pair case in 3-D, etc.

[00146] A key problem is how to apply spatial filtering to such a combination of paired 1-D DOA estimates. In this case, a beamformer/null beamformer (BFNF) as shown in Figure 43 may be applied by augmenting the steering vector for each pair. In this figure, A^H denotes the conjugate transpose of A, x denotes the microphone channels, and y denotes the spatially filtered channels. Using a pseudo-inverse operation $A^+ = (A^H A)^{-1} A^H$ as shown in Figure 43 allows the use of a non-square matrix. For a three-microphone case (i.e., two microphone pairs) as illustrated in Figure 45, for example, the number of rows 2*2=4 instead of 3, such that the additional row makes the matrix non-square.

[00147] As the approach shown in Figure 43 is based on robust 1-D DOA estimation, complete knowledge of the microphone geometry is not required, and DOA estimation using all microphones at the same time is also not required. Such an approach is well-suited for use with anglogram-based DOA estimation as described herein, although any other 1-D DOA estimation method can also be used. Figure 44 shows an example of the BFNF as shown in Figure 43 which also includes a normalization factor to prevent an ill-conditioned inversion at the spatial aliasing frequency.

[00148] Figure 46 shows an example of a pair-wise (PW) normalized MVDR (minimum variance distortionless response) BFNF, in which the manner in which the steering vector (array manifold vector) is obtained differs from the conventional approach. In this case, a common channel is eliminated due to sharing of a microphone between the two pairs. The noise coherence matrix Γ may be obtained either by measurement or by theoretical calculation using a sinc function. It is noted that the examples of Figures 43, 44, and 46 may be generalized to an arbitrary number of sources N such that N <= M, where M is the number of microphones.

[00149] Figure 47 shows another example that may be used if the matrix A^HA is not ill-conditioned, which may be determined using a condition number or determinant of the matrix. If the matrix is ill-conditioned, it may be desirable to bypass one microphone signal for that frequency bin for use as the source channel, while continuing to apply the method to spatially filter other frequency bins in which the matrix A^HA is not ill-conditioned. This option saves computation for calculating a denominator for normalization. The methods in Figures 43-47 demonstrate BFNF techniques that may be applied independently at each frequency bin. The steering vectors are constructed using the DOA estimates for each frequency and microphone pair as described herein. For example, each element of the steering vector for pair p and source n for DOA θ_i, frequency f, and microphone number m (1 or 2) may be calculated as

$$d_{p,m}^{n} = \exp\left(\frac{-j\omega f_s(m-1)l_p}{c}\cos\theta_i\right),$$

where l_p indicates the distance between the microphones of pair p, ω indicates the frequency bin number, and f_s indicates the sampling frequency. Figure 48 shows examples of steering vectors for an array as shown in Figure 45.

[00150] A PWBFNF scheme may be used for suppressing direct path of interferers up to the available degrees of freedom (instantaneous suppression without smooth trajectory assumption, additional noise-suppression gain using directional masking, additional noise-suppression gain using bandwidth extension). Single-channel post-processing of quadrant framework may be used for stationary noise and noise-reference handling.

[00151] It may be desirable to obtain instantaneous suppression but also to provide minimization of artifacts such as musical noise. It may be desirable to maximally use the available degrees of freedom for BFNF. One DOA may be fixed across all frequencies, or a slightly mismatched alignment across frequencies may be permitted. Only the current frame may be used, or a feed-forward network may be implemented.

The BFNF may be set for all frequencies in the range up to the Nyquist rate (e.g., except ill-conditioned frequencies). A natural masking approach may be used (e.g., to obtain a smooth natural seamless transition of aggressiveness).

[00152] Figure 49 shows a flowchart for one example of an integrated method as described herein. This method includes an inventory matching task for phase delay estimation, a variance calculation task to obtain DOA error variance values, a dimension-matching and/or pair-selection task, and a task to map DOA error variance for the selected DOA candidate to a source activity likelihood estimate. The pair-wise DOA estimation results may also be used to track one or more active speakers, to perform a pair-wise spatial filtering operation, and or to perform time- and/or frequency-selective masking. The activity likelihood estimation and/or spatial filtering operation may also be used to obtain a noise estimate to support a single-channel noise suppression operation.

[00153] The methods and apparatus disclosed herein may be applied generally in any transceiving and/or audio sensing application, especially mobile or otherwise portable instances of such applications. For example, the range of configurations disclosed herein includes communications devices that reside in a wireless telephony communication system configured to employ a code-division multiple-access (CDMA) over-the-air interface. Nevertheless, it would be understood by those skilled in the art that a method and apparatus having features as described herein may reside in any of the various communication systems employing a wide range of technologies known to those of skill in the art, such as systems employing Voice over IP (VoIP) over wired and/or wireless (e.g., CDMA, TDMA, FDMA, and/or TD-SCDMA) transmission channels.

[00154] It is expressly contemplated and hereby disclosed that communications devices disclosed herein may be adapted for use in networks that are packet-switched (for example, wired and/or wireless networks arranged to carry audio transmissions according to protocols such as VoIP) and/or circuit-switched. It is also expressly contemplated and hereby disclosed that communications devices disclosed herein may be adapted for use in narrowband coding systems (e.g., systems that encode an audio frequency range of about four or five kilohertz) and/or for use in wideband coding systems (e.g., systems that encode audio frequencies greater than five kilohertz), including whole-band wideband coding systems and split-band wideband coding systems.

[00155] Examples of codecs that may be used with, or adapted for use with, transmitters and/or receivers of communications devices as described herein include the Enhanced Variable Rate Codec, as described in the Third Generation Partnership Project 2 (3GPP2) document C.S0014-C, v1.0, entitled "Enhanced Variable Rate Codec, Speech Service Options 3, 68, and 70 for Wideband Spread Spectrum Digital Systems," February 2007 (available online at www-dot-3gpp-dot-org); the Selectable Mode Vocoder speech codec, as described in the 3GPP2 document C.S0030-0, v3.0, entitled "Selectable Mode Vocoder (SMV) Service Option for Wideband Spread Spectrum Communication Systems," January 2004 (available online at www-dot-3gpp-dot-org); the Adaptive Multi Rate (AMR) speech codec, as described in the document ETSI TS 126 092 V6.0.0 (European Telecommunications Standards Institute (ETSI), Sophia Antipolis Cedex, FR, December 2004); and the AMR Wideband speech codec, as described in the document ETSI TS 126 192 V6.0.0 (ETSI, December 2004). Such a codec may be used, for example, to recover the reproduced audio signal from a received wireless communications signal.

[00156] The presentation of the described configurations is provided to enable any person skilled in the art to make or use the methods and other structures disclosed herein. The flowcharts, block diagrams, and other structures shown and described herein are examples only, and other variants of these structures are also within the scope of the disclosure. Various modifications to these configurations are possible, and the generic principles presented herein may be applied to other configurations as well. Thus, the present disclosure is not intended to be limited to the configurations shown above but rather is to be accorded the widest scope consistent with the principles and novel features disclosed in any fashion herein, including in the attached claims as filed, which form a part of the original disclosure.

[00157] Those of skill in the art will understand that information and signals may be represented using any of a variety of different technologies and techniques. For example, data, instructions, commands, information, signals, bits, and symbols that may be referenced throughout the above description may be represented by voltages, currents, electromagnetic waves, magnetic fields or particles, optical fields or particles, or any combination thereof.

[00158] Important design requirements for implementation of a configuration as disclosed herein may include minimizing processing delay and/or computational complexity (typically measured in millions of instructions per second or MIPS), especially for computation-intensive applications, such as playback of compressed audio or audiovisual information (e.g., a file or stream encoded according to a compression format, such as one of the examples identified herein) or applications for wideband communications (e.g., voice communications at sampling rates higher than eight kilohertz, such as 12, 16, 32, 44.1, 48, or 192 kHz).

[00159] An apparatus as disclosed herein (e.g., any device configured to perform a technique as described herein) may be implemented in any combination of hardware with software, and/or with firmware, that is deemed suitable for the intended application. For example, the elements of such an apparatus may be fabricated as electronic and/or optical devices residing, for example, on the same chip or among two or more chips in a chipset. One example of such a device is a fixed or programmable array of logic elements, such as transistors or logic gates, and any of these elements may be implemented as one or more such arrays. Any two or more, or even all, of these elements may be implemented within the same array or arrays. Such an array or arrays may be implemented within one or more chips (for example, within a chipset including two or more chips).

[00160] One or more elements of the various implementations of the apparatus disclosed herein may be implemented in whole or in part as one or more sets of instructions arranged to execute on one or more fixed or programmable arrays of logic elements, such as microprocessors, embedded processors, IP cores, digital signal processors, FPGAs (field-programmable gate arrays), ASSPs (application-specific standard products), and ASICs (application-specific integrated circuits). Any of the various elements of an implementation of an apparatus as disclosed herein may also be embodied as one or more computers (e.g., machines including one or more arrays programmed to execute one or more sets or sequences of instructions, also called "processors"), and any two or more, or even all, of these elements may be implemented within the same such computer or computers.

[00161] A processor or other means for processing as disclosed herein may be fabricated as one or more electronic and/or optical devices residing, for example, on the same chip or among two or more chips in a chipset. One example of such a device is a fixed or programmable array of logic elements, such as transistors or logic gates, and any of these elements may be implemented as one or more such arrays. Such an array or arrays may be implemented within one or more chips (for example, within a chipset including two or more chips). Examples of such arrays include fixed or programmable arrays of logic elements, such as microprocessors, embedded processors, IP cores, DSPs, FPGAs, ASSPs, and ASICs. A processor or other means for processing as disclosed herein may also be embodied as one or more computers (e.g., machines including one or more arrays programmed to execute one or more sets or sequences of instructions) or other processors. It is possible for a processor as described herein to be used to perform tasks or execute other sets of instructions that are not directly related to a procedure of an implementation described herein, such as a task relating to another operation of a device or system in which the processor is embedded (e.g., an audio sensing device). It is also possible for part of a method as disclosed herein to be performed by a processor of the audio sensing device and for another part of the method to be performed under the control of one or more other processors.

[00162] Those of skill will appreciate that the various illustrative modules, logical blocks, circuits, and tests and other operations described in connection with the configurations disclosed herein may be implemented as electronic hardware, computer software, or combinations of both. Such modules, logical blocks, circuits, and operations may be implemented or performed with a general purpose processor, a digital signal processor (DSP), an ASIC or ASSP, an FPGA or other programmable logic device, discrete gate or transistor logic, discrete hardware components, or any combination thereof designed to produce the configuration as disclosed herein. For example, such a configuration may be implemented at least in part as a hard-wired circuit, as a circuit configuration fabricated into an application-specific integrated circuit, or as a firmware program loaded into non-volatile storage or a software program loaded from or into a data storage medium as machine-readable code, such code being instructions executable by an array of logic elements such as a general purpose processor or other digital signal processing unit. A general purpose processor may be a microprocessor, but in the alternative, the processor may be any conventional processor, controller, microcontroller, or state machine. A processor may also be implemented as a combination of computing devices, e.g., a combination of a DSP and a microprocessor, a plurality of microprocessors, one or more microprocessors in conjunction with a DSP core, or any other such configuration. A software module may reside in a non-transitory storage medium such as RAM (random-access memory), ROM (read-only memory), nonvolatile RAM (NVRAM) such as flash RAM, erasable programmable ROM (EPROM), electrically erasable programmable ROM (EEPROM), registers, hard disk, a removable disk, or a CD-ROM; or in any other form of storage medium known in the art. An illustrative storage medium is coupled to the processor such the processor can read information from, and write information to, the storage medium. In the alternative, the storage medium may be integral to the processor. The processor and the storage medium may reside in an ASIC. The ASIC may reside in a user terminal. In the alternative, the processor and the storage medium may reside as discrete components in a user terminal.

[00163] It is noted that the various methods disclosed herein may be performed by an array of logic elements such as a processor, and that the various elements of an apparatus as described herein may be implemented as modules designed to execute on such an array. As used herein, the term "module" or "sub-module" can refer to any method, apparatus, device, unit or computer-readable data storage medium that includes computer instructions (e.g., logical expressions) in software, hardware or firmware form. It is to be understood that multiple modules or systems can be combined into one module or system and one module or system can be separated into multiple modules or systems to perform the same functions. When implemented in software or other computer-executable instructions, the elements of a process are essentially the code segments to perform the related tasks, such as with routines, programs, objects, components, data structures, and the like. The term “software” should be understood to include source code, assembly language code, machine code, binary code, firmware, macrocode, microcode, any one or more sets or sequences of instructions executable by an array of logic elements, and any combination of such examples. The program or code segments can be stored in a processor readable medium or transmitted by a computer data signal embodied in a carrier wave over a transmission medium or communication link.

[00164] Each of the tasks of the methods described herein may be embodied directly in hardware, in a software module executed by a processor, or in a combination of the two. In a typical application of an implementation of a method as disclosed herein, an array of logic elements (e.g., logic gates) is configured to perform one, more than one, or even all of the various tasks of the method. One or more (possibly all) of the tasks may also be implemented as code (e.g., one or more sets of instructions), embodied in a computer program product (e.g., one or more data storage media such as disks, flash or other nonvolatile memory cards, semiconductor memory chips, etc.), that is readable and/or executable by a machine (e.g., a computer) including an array of logic elements (e.g., a processor, microprocessor, microcontroller, or other finite state machine). The tasks of an implementation of a method as disclosed herein may also be performed by more than one such array or machine. In these or other implementations, the tasks may be performed within a device for wireless communications such as a cellular telephone or other device having such communications capability. Such a device may be configured to communicate with circuit-switched and/or packet-switched networks (e.g., using one or more protocols such as VoIP). For example, such a device may include RF circuitry configured to receive and/or transmit encoded frames.

[00165] It is expressly disclosed that the various methods disclosed herein may be performed by a portable communications device such as a handset, headset, or portable digital assistant (PDA), and that the various apparatus described herein may be included within such a device.

[00166] In one or more exemplary embodiments, the operations described herein may be implemented in hardware, software, firmware, or any combination thereof. If implemented in software, such operations may be stored on or transmitted over a computer-readable medium as one or more instructions or code. The term "computer-readable media" includes both computer-readable storage media and communication (e.g., transmission) media. By way of example, and not limitation, computer-readable storage media can comprise an array of storage elements, such as semiconductor memory (which may include without limitation dynamic or static RAM, ROM, EEPROM, and/or flash RAM), or ferroelectric, magnetoresistive, ovonic, polymeric, or phase-change memory; CD-ROM or other optical disk storage; and/or magnetic disk storage or other magnetic storage devices. Such storage media may store information in the form of instructions or data structures that can be accessed by a computer. Communication media can comprise any medium that can be used to carry desired program code in the form of instructions or data structures and that can be accessed by a computer, including any medium that facilitates transfer of a computer program from one place to another. Also, any connection is properly termed a computer-readable medium. For example, if the software is transmitted from a website, server, or other remote source using a coaxial cable, fiber optic cable, twisted pair, digital subscriber line (DSL), or wireless technology such as infrared, radio, and/or microwave, then the coaxial cable, fiber optic cable, twisted pair, DSL, or wireless technology such as infrared, radio, and/or microwave are included in the definition of medium. Disk and disc, as used herein, includes compact disc (CD), laser disc, optical disc, digital versatile disc (DVD), floppy disk and Blu-ray DiscTM (Blu-Ray Disc Association, Universal City, CA), where disks usually reproduce data magnetically, while discs reproduce data optically with lasers. Combinations of the above should also be included within the scope of computer-readable media.

[00167] An acoustic signal processing apparatus as described herein may be incorporated into an electronic device that accepts speech input in order to control certain operations, or may otherwise benefit from separation of desired noises from background noises, such as communications devices. Many applications may benefit from enhancing or separating clear desired sound from background sounds originating from multiple directions. Such applications may include human-machine interfaces in electronic or computing devices which incorporate capabilities such as voice recognition and detection, speech enhancement and separation, voice-activated control, and the like. It may be desirable to implement such an acoustic signal processing apparatus to be suitable in devices that only provide limited processing capabilities.

[00168] It is possible for one or more elements of an implementation of an apparatus as described herein to be used to perform tasks or execute other sets of instructions that are not directly related to an operation of the apparatus, such as a task relating to another operation of a device or system in which the apparatus is embedded. It is also possible for one or more elements of an implementation of such an apparatus to have structure in common (e.g., a processor used to execute portions of code corresponding to different elements at different times, a set of instructions executed to perform tasks corresponding to different elements at different times, or an arrangement of electronic and/or optical devices performing operations for different elements at different times).

[00169] The previous description of the disclosure is provided to enable any person skilled in the art to make or use the disclosure. Various modifications to the disclosure will be readily apparent to those skilled in the art, and the generic principles defined herein may be applied to other variations without departing from the scope of the disclosure. Thus, the disclosure is not intended to be limited to the examples and designs described herein but is to be accorded the widest scope consistent with the principles and novel features disclosed herein.

[00170] Although exemplary implementations may refer to utilizing aspects of the presently disclosed subject matter in the context of one or more stand-alone computer systems, the subject matter is not so limited, but rather may be implemented in connection with any computing environment, such as a network or distributed computing environment. Still further, aspects of the presently disclosed subject matter may be implemented in or across a plurality of processing chips or devices, and storage may similarly be effected across a plurality of devices. Such devices might include PCs, network servers, and handheld devices, for example.

[00171] Although the subject matter has been described in language specific to structural features and/or methodological acts, it is to be understood that the subject matter defined in the appended claims is not necessarily limited to the specific features or acts described above. Rather, the specific features and acts described above are disclosed as example forms of implementing the claims.

WHAT IS CLAIMED:

1. A system which tracks a social interaction between a plurality of participants, comprising:

a fixed beamformer that is adapted to output a first spatially filtered output, and configured to receive a plurality of second spatially filtered outputs from a plurality of steerable beamformers, each steerable beamformer outputting a respective one of the second spatially filtered outputs and associated with a different one of the participants; and

a processor capable of determining a similarity between the first spatially filtered output and each of the second spatially filtered outputs, and capable of determining the social interaction between the participants based on the similarity between the first spatially filtered output and each of the second spatially filtered outputs.

2. The system of claim 1, wherein the fixed beamformer comprises a fixed microphone array, and each of the steerable beamformers comprises a steerable microphone array.

3. The system of claim 1, wherein the fixed beamformer and the processor are comprised within a mobile device.

4. The system of claim 1, wherein the fixed beamformer and the processor are comprised within one from among the group comprising a handset, a laptop, a tablet, a computer, and a netbook.

5. The system of claim 1, wherein each of the plurality of steerable beamformers is comprised within a respective device, and wherein each respective device is capable of being associated with a different one of the participants.

6. The system of claim 5, wherein each respective device comprises a headset worn by the associated participant.

7. The system of claim 1, further comprising a user interface which is capable of displaying the social interaction between the participants.

8. The system of claim 7, wherein a user interface display is capable of graphically displaying the plurality of participants at once.

9. The system of claim 8, wherein the user interface display is capable of zooming in on one of the participants via the user interface to provide an enhanced speech of the zoomed in participant.

10. The system of claim 1, wherein the first spatially filtered output of the fixed beamformer is refined based on at least one of the second spatially filtered outputs of the plurality of steerable beamformers.

11. The system of claim 1, wherein the processor is adapted to compare (1) the similarity between the first spatially filtered output of the fixed beamformer and the at least one of the second spatially filtered outputs of the plurality of steerable beamformers with (2) the similarity between the first spatially filtered output of a fixed beamformer having a moved look direction and the at least one of the second spatially filtered outputs of the plurality of steerable beamformers.

12. The system of claim 1, further comprising a mobile device capable of tracking the social interaction between participants based on audio beamforming.

13. The system of claim 1, wherein the processor is configured to perform a correlation between the first spatially filtered output of the fixed beamformer and a selected one of the second spatially filtered outputs of the steerable beamformers.

14. The system of claim 13, wherein the fixed beamformer is comprised within a first mobile device, and a selected steerable beamformer is comprised within a second mobile device that is different from the first mobile device.

15. The system of claim 1, wherein the similarity is determined using one of: a correlation, a least square fit with allowable time adjustment in the time domain or the frequency domain, a feature based approach using linear prediction coding (LPC) or mel-frequency cepstral coefficients (MFCC), or a higher order based approach using cross-cumulant, an empirical Kullback-Leibler divergence, or an Itakura-Saito distance.

16. The system of claim 1, wherein the processor is further configured to determine a location of at least one of the participants.

17. A system for determining a similarity between an output of a fixed microphone array and an output of a steerable microphone array, comprising:

a processor that is configured to receive a first spatially filtered output from the fixed microphone array and a second spatially filtered output from the steerable microphone array, and is further configured to compare the first spatially filtered output and the second spatially filtered output to determine the similarity between the output of the fixed microphone array and the output of the steerable microphone array; and

an output device that is configured to output the similarity.

18. The system of claim 17, wherein each spatially filtered output comprises a beamformed output.

19. The system of claim 17, wherein the processor is further configured to repeat the receiving and comparing a plurality of times, once for each of a plurality of steerable microphone arrays.

20. The system of claim 17, wherein the processor and the output device are comprised within one from among the group comprising a handset, a laptop, a tablet, a computer, and a netbook.

21. The system of claim 17, wherein the processor is further configured to:

estimate a direction of arrival around the fixed microphone array;

determine an active speaker using the direction of arrival, and separate an output of the active speaker using the direction of arrival; and

determine the similarity between the output of the steerable microphone array and the output of the fixed microphone array using the first spatially filtered output, the second spatially filtered output, and the output of the active speaker.

22. The system of claim 21, wherein estimating the direction of arrival may be performed in three dimensions (3D).

23. The system of claim 21, wherein the second spatially filtered output is in an active speaker look direction.

24. The system of claim 21, wherein the second spatially filtered output is generated by fixed broadside beamforming from an active noise control (ANC) headset.

25. The system of claim 21, wherein the processor is further configured to:

receive a plurality of spatially filtered outputs from a plurality of steerable microphone arrays, each steerable microphone array corresponding to a different active speaker;

identify the active speakers, and separate the outputs of the active speakers using the direction of arrival; and

determine the similarities between the outputs of the steerable microphone arrays and the separated outputs, corresponding to the active speakers, of the fixed microphone array.

26. The system of claim 25, wherein the fixed microphone array is configured to provide a number of active speakers and a separated speech signal for each active speaker, and the steerable microphone array provides a look direction of each active speaker, and determining the similarities comprises:

for each active speaker;

finding a maximum peak of a cross-correlation equation using the separated speech signal for the active speaker and the look direction of the active speaker, and

determining an angle of strong correlation for using the maximum peak, wherein the angle of strong correlation corresponds to an angle between the fixed microphone array and the steerable microphone array of the active speaker.

27. The system of claim 25, wherein determining the similarities uses one of: a correlation, a least square fit with allowable time adjustment in the time domain or the frequency domain, a feature based approach using linear prediction coding (LPC) or mel-frequency cepstral coefficients (MFCC), or a higher order based approach using cross-cumulant, an empirical Kullback-Leibler divergence, or an Itakura-Saito distance.

28. A method for tracking a social interaction between a plurality of participants, comprising:

outputting, from a fixed beamformer, a first spatially filtered output;

outputting, from a plurality of steerable beamformers, a plurality of second spatially filtered outputs, each steerable beamformer outputting a respective one of the second spatially filtered outputs and associated with a different one of the participants;

determining a similarity between the first spatially filtered output and each of the second spatially filtered outputs; and

determining, utilizing a processor, the social interaction between the participants based on the similarity between the first spatially filtered output and each of the second spatially filtered outputs.

29. The method of claim 28, wherein the fixed beamformer comprises a fixed microphone array, and each of the steerable beamformers comprises a steerable microphone array.

30. The method of claim 28, wherein the fixed beamformer and the processor are comprised within a mobile device.

31. The method of claim 28, wherein the fixed beamformer and the processor are comprised within one from among the group comprising a handset, a laptop, a tablet, a computer, and a netbook.

32. The method of claim 28, wherein each of the plurality of steerable beamformers is comprised within a respective device, and wherein each respective device is capable of being associated with a different one of the participants.

33. The method of claim 32, wherein each respective device comprises a headset worn by the associated participant.

34. The method of claim 28, further comprising displaying the social interaction between the participants.

35. The method of claim 34, wherein the displaying further comprises graphically displaying the plurality of participants at once.

36. The method of claim 35, wherein the displaying further comprises zooming in on one of the participants via a user interface to provide an enhanced speech of the zoomed in participant.

37. The method of claim 28, further comprising refining the first spatially filtered output of the fixed beamformer based on at least one of the second spatially filtered outputs of the plurality of steerable beamformers.

38. The method of claim 28, further comprising comparing (1) the similarity between the first spatially filtered output of the fixed beamformer and the at least one of the second spatially filtered outputs of the plurality of steerable beamformers with (2) the similarity between the first spatially filtered output of a fixed beamformer having a moved look direction and the at least one of the second spatially filtered outputs of the plurality of steerable beamformers.

39. The method of claim 28, further comprising tracking the social interaction between participants based on audio beamforming.

40. The method of claim 28, further comprising performing a correlation between the first spatially filtered output of the fixed beamformer and a selected one of the second spatially filtered outputs of the steerable beamformers.

41. The method of claim 40, wherein the fixed beamformer is comprised within a first mobile device, and a selected steerable beamformer is comprised within a second mobile device that is different from the first mobile device.

42. The method of claim 28, wherein the similarity is determined using one of: a least square fit with allowable time adjustment in the time domain or the frequency domain, a feature based approach using linear prediction coding (LPC) or mel-frequency cepstral coefficients (MFCC), or a higher order based approach using cross-cumulant, an empirical Kullback-Leibler divergence, or an Itakura-Saito distance.

43. The method of claim 28, further comprising determining a location of at least one of the participants.

44. A method for determining a similarity between an output of a fixed microphone array and an output of a steerable microphone array, comprising:

receiving a first spatially filtered output from the fixed microphone array;

receiving a second spatially filtered output from a steerable microphone array;

comparing the first spatially filtered output and the second spatially filtered output;

determining the similarity between the output of the fixed microphone array and the output of the steerable microphone array based on the comparison; and

outputting the similarity.

45. The method of claim 44, wherein each spatially filtered output comprises a beamformed output.

46. The method of claim 44, further comprising repeating the receiving and comparing a plurality of times, once for each of a plurality of steerable microphone arrays.

47. The method of claim 44, wherein the fixed microphone array is comprised within one from among the group comprising a handset, a laptop, a tablet, a computer, and a netbook.

48. The method of claim 44, further comprising:

estimating a direction of arrival around the fixed microphone array;

determining an active speaker using the direction of arrival;

separating an output of the active speaker using the direction of arrival; and

determining the similarity between the output of the steerable microphone array and the output of the fixed microphone array using the first spatially filtered output, the second spatially filtered output, and the output of the active speaker.

49. The method of claim 48, wherein estimating the direction of arrival may be performed in three dimensions (3D).

50. The method of claim 48, wherein the second spatially filtered output is in an active speaker look direction.

51. The method of claim 48, wherein the second spatially filtered output is generated by fixed broadside beamforming from an active noise control (ANC) headset.

52. The method of claim 48, further comprising:

receiving a plurality of spatially filtered outputs from a plurality of steerable microphone arrays, each steerable microphone array corresponding to a different active speaker;

identifying the active speaker, and separate the outputs of the active speaker using the direction of arrival; and

determining similarities between the outputs of the steerable microphone arrays and the separated outputs, corresponding to the active speakers, of the fixed microphone array.

53. The method of claim 52, further comprising providing a number of active speakers and a separated speech signal for each active speaker and providing a look direction of each active speaker, wherein determining the similarities comprises, for each active speaker:

finding a maximum peak of a cross-correlation equation using the separated speech signal for the active speaker and the look direction of the active speaker; and

determining an angle of strong correlation for using the maximum peak, wherein the angle of strong correlation corresponds to an angle between the fixed microphone array and the steerable microphone array of the active speaker.

54. The method of claim 52, wherein determining the similarities uses one of: a correlation, a least square fit with allowable time adjustment in the time domain or the frequency domain, a feature based approach using linear prediction coding (LPC) or mel-frequency cepstral coefficients (MFCC), or a higher order based approach using cross-cumulant, an empirical Kullback-Leibler divergence, or an Itakura-Saito distance.

55. An apparatus for tracking a social interaction between a plurality of participants, comprising:

means for outputting a first spatially filtered output;

means for outputting a plurality of second spatially filtered outputs, each of the second spatially filtered outputs associated with a different one of the participants;

means for determining a similarity between the first spatially filtered output and each of the second spatially filtered outputs; and means for determining the social interaction between the participants based on the similarity between the first spatially filtered output and each of the second spatially filtered outputs.

56. The apparatus of claim 55, wherein the means are comprised within a mobile device.

57. The apparatus of claim 55, wherein the means are comprised within one from among the group comprising a handset, a laptop, a tablet, a computer, and a netbook.

58. The apparatus of claim 55, further comprising means for displaying the social interaction between the participants.

59. The apparatus of claim 58, further comprising means for graphically displaying the plurality of participants at once.

60. The apparatus of claim 59, further comprising means for zooming in on one of the participants via a user interface to provide an enhanced speech of the zoomed in participant.

61. The apparatus of claim 55, further comprising means for refining the first spatially filtered output based on at least one of the second spatially filtered outputs.

62. The apparatus of claim 55, further comprising means for comparing (1) the similarity between the first spatially filtered output and the at least one of the second spatially filtered outputs with (2) the similarity between a first spatially filtered output having a moved look direction and the at least one of the second spatially filtered outputs.

63. The apparatus of claim 55, further comprising means for tracking the social interaction between participants based on audio beamforming.

64. The apparatus of claim 55, further comprising means for performing a correlation between the first spatially filtered output and a selected one of the second spatially filtered outputs.

65. The apparatus of claim 55, wherein the similarity is determined using one of: a correlation, a least square fit with allowable time adjustment in the time domain or the frequency domain, a feature based approach using linear prediction coding (LPC) or mel-frequency cepstral coefficients (MFCC), or a higher order based approach using cross-cumulant, an empirical Kullback-Leibler divergence, or an Itakura-Saito distance.

66. The apparatus of claim 55, further comprising means for determining a location of at least one of the participants.

67. An apparatus for determining a similarity between an output of a fixed microphone array and an output of a steerable microphone array, comprising:

means for receiving a first spatially filtered output from the fixed microphone array;

means for receiving a second spatially filtered output from the steerable microphone array;

means for comparing the first spatially filtered output and the second spatially filtered output ;

means for determining the similarity between the output of the fixed microphone array and the output of the steerable microphone array based on the comparison; and

means for outputting the similarity.

68. The apparatus of claim 67, wherein each spatially filtered output comprises a beamformed output.

69. The apparatus of claim 67, further comprising means for repeating the receiving and comparing a plurality of times, once for each of a plurality of steerable microphone arrays.

70. The apparatus of claim 67, wherein the means for receiving are comprised within one from among the group comprising a handset, a laptop, a tablet, a computer, and a netbook.

71. The apparatus of claim 67, further comprising:

means for estimating a direction of arrival around the fixed microphone array;

means for determining an active speaker using the direction of arrival;

means for separating an output of the active speaker using the direction of arrival; and

means for determining the similarity between the output of the steerable microphone array and the output of the fixed microphone array using the first spatially filtered output, the second spatially filtered output, and the output of the active speaker.

72. The apparatus of claim 71, wherein the means for estimating the direction of arrival is capable of performing the estimate in three dimensions (3D).

73. The apparatus of claim 71, wherein the second spatially filtered output is in an active speaker look direction.

74. The apparatus of claim 71, wherein the second spatially filtered output is generated by fixed broadside beamforming from an active noise control (ANC) headset.

75. The apparatus of claim 71, further comprising:

means for receiving a plurality of spatially filtered outputs from a plurality of steerable microphone arrays, each steerable microphone array corresponding to a different active speaker;

means for identifying the active speakers and separating the outputs of the active speakers using the direction of arrival; and

means for determining the similarities between the outputs of the steerable microphone arrays and the separated outputs, corresponding to the active speakers, of the fixed microphone array.

76. The apparatus of claim 75, further comprising means for providing a number of active speakers and a separated speech signal for each of the active speakers and means for providing a look direction of each of the active speakers, wherein determining the similarities comprises, for each of the active speakers:

finding a maximum peak of a cross-correlation equation using the separated speech signal for the active speaker and the look direction of the active speaker; and

determining an angle of strong correlation for using the maximum peak, wherein the angle of strong correlation corresponds to an angle between the fixed microphone array and the steerable microphone array of the active speaker.

77. The apparatus of claim 75, wherein determining the similarities uses one of: a correlation, a least square fit with allowable time adjustment in the time domain or the frequency domain, a feature based approach using linear prediction coding (LPC) or mel-frequency cepstral coefficients (MFCC), or a higher order based approach using cross-cumulant, an empirical Kullback-Leibler divergence, or an Itakura-Saito distance.

78. A non-transitory computer-readable medium comprising computer-readable instructions for causing a processor to:

receive, from a fixed beamformer, a first spatially filtered output;

receive, from a plurality of steerable beamformers, a plurality of second spatially filtered outputs, each steerable beamformer outputting a respective one of the second spatially filtered outputs and associated with a different one of the participants;

determine a similarity between the first spatially filtered output and each of the second spatially filtered outputs; and

determine, utilizing a processor, a social interaction between the participants based on the similarity between the first spatially filtered output and each of the second spatially filtered outputs.

79. The computer-readable medium of claim 78, wherein the fixed beamformer comprises a fixed microphone array, and each of the steerable beamformers comprises a steerable microphone array.

80. The computer-readable medium of claim 78, wherein the fixed beamformer and the processor are comprised within a mobile device.

81. The computer-readable medium of claim 78, wherein the fixed beamformer and the processor are comprised within one from among the group comprising a handset, a laptop, a tablet, a computer, and a netbook.

82. The computer-readable medium of claim 78, wherein each of the plurality of steerable beamformers is comprised within a respective device, and wherein each respective device is capable of being associated with a different one of the participants.

83. The computer-readable medium of claim 82, wherein each respective device comprises a headset worn by the associated participant.

84. The computer-readable medium of claim 78, further comprising instructions for causing the processor to display the social interaction between the participants.

85. The computer-readable medium of claim 84, further comprising instructions for causing the processor to graphically display the plurality of participants at once.

86. The computer-readable medium of claim 85, further comprising instructions for causing the processor to zoom in on one of the participants via a user interface to provide an enhanced speech of the zoomed in participant.

87. The computer-readable medium of claim 78, further comprising instructions for causing the processor to refine the first spatially filtered output of the fixed beamformer based on at least one of the second spatially filtered outputs of the plurality of steerable beamformers.

88. The computer-readable medium of claim 78, further comprising instructions for causing the processor to compare (1) the similarity between the first spatially filtered output of the fixed beamformer and the at least one of the second spatially filtered outputs of the plurality of steerable beamformers with (2) the similarity between the first spatially filtered output of a fixed beamformer having a moved look direction and the at least one of the second spatially filtered outputs of the plurality of steerable beamformers.

89. The computer-readable medium of claim 78, further comprising instructions for causing the processor to track the social interaction between participants based on audio beamforming.

90. The computer-readable medium of claim 78, further comprising instructions for causing the processor to perform a correlation between the first spatially filtered output of the fixed beamformer and a selected one of the second spatially filtered outputs of the steerable beamformers.

91. The computer-readable medium of claim 90, wherein the fixed beamformer is comprised within a first mobile device, and a selected steerable beamformer is comprised within a second mobile device that is different from the first mobile device.

92. The computer-readable medium of claim 78, wherein the similarity is determined using one of: a least square fit with allowable time adjustment in the time domain or the frequency domain, a feature based approach using linear prediction coding (LPC) or mel-frequency cepstral coefficients (MFCC), or a higher order based approach using cross-cumulant, an empirical Kullback-Leibler divergence, or an Itakura-Saito distance.

93. The computer-readable medium of claim 78, further comprising instructions for causing the processor to determine a location of at least one of the participants.

94. A non-transitory computer-readable medium comprising computer-readable instructions for causing a processor to:

receive a first spatially filtered output from a fixed microphone array;

receive a second spatially filtered output from a steerable microphone array;

compare the first spatially filtered output and the second spatially filtered output;

determine a similarity between the output of the fixed microphone array and the output of the steerable microphone array based on the comparison; and

output the similarity.

95. The computer-readable medium of claim 94, wherein each spatially filtered output comprises a beamformed output.

96. The computer-readable medium of claim 94, further comprising instructions for causing the processor to repeat the receive and compare a plurality of times, once for each of a plurality of steerable microphone arrays.

97. The computer-readable medium of claim 94, wherein the processor is comprised within one from among the group comprising a handset, a laptop, a tablet, a computer, and a netbook.

98. The computer-readable medium of claim 94, further comprising instructions for causing the processor to:

estimate a direction of arrival around the fixed microphone array;

determine an active speaker using the direction of arrival;

separate an output of the active speaker using the direction of arrival; and determine the similarity between the output of the steerable microphone array and the output of the fixed microphone array using the first spatially filtered output, the second spatially filtered output, and the output of the active speaker.

99. The computer-readable medium of claim 98, further comprising instructions for causing the processor to estimating the direction of arrival in three dimensions (3D).

100. The computer-readable medium of claim 98, wherein the second spatially filtered output is in an active speaker look direction.

101. The computer-readable medium of claim 98, wherein the second spatially filtered output is generated by fixed broadside beamforming from an active noise control (ANC) headset.

102. The computer-readable medium of claim 98, further comprising instructions for causing the processor to:

receive a plurality of spatially filtered outputs from a plurality of steerable microphone arrays, each steerable microphone array corresponding to a different active speaker;

identify the active speakers, and separate the outputs of the active speakers using the direction of arrival; and

determine the similarities between the outputs of the steerable microphone arrays and the separated outputs, corresponding to the active speakers, of the fixed microphone array.

103. The computer-readable medium of claim 102, further comprising instructions for causing the processor to provide a number of active speakers and a separated speech signal for each of the active speakers and providing a look direction of each of the active speakers, wherein determining the similarities comprises, for each of the active speakers comprises:

finding a maximum peak of a cross-correlation equation using the separated speech signal for the active speaker and the look direction of the active speaker; and

determining an angle of strong correlation for using the maximum peak, wherein the angle of strong correlation corresponds to an angle between the fixed microphone array and the steerable microphone array of the active speaker.

104. The computer-readable medium of claim 102, wherein determining the similarities uses one of: a correlation, a least square fit with allowable time adjustment in the time domain or the frequency domain, a feature based approach using linear prediction coding (LPC) or mel-frequency cepstral coefficients (MFCC), or a higher order based approach using cross-cumulant, an empirical Kullback-Leibler divergence, or an Itakura-Saito distance.

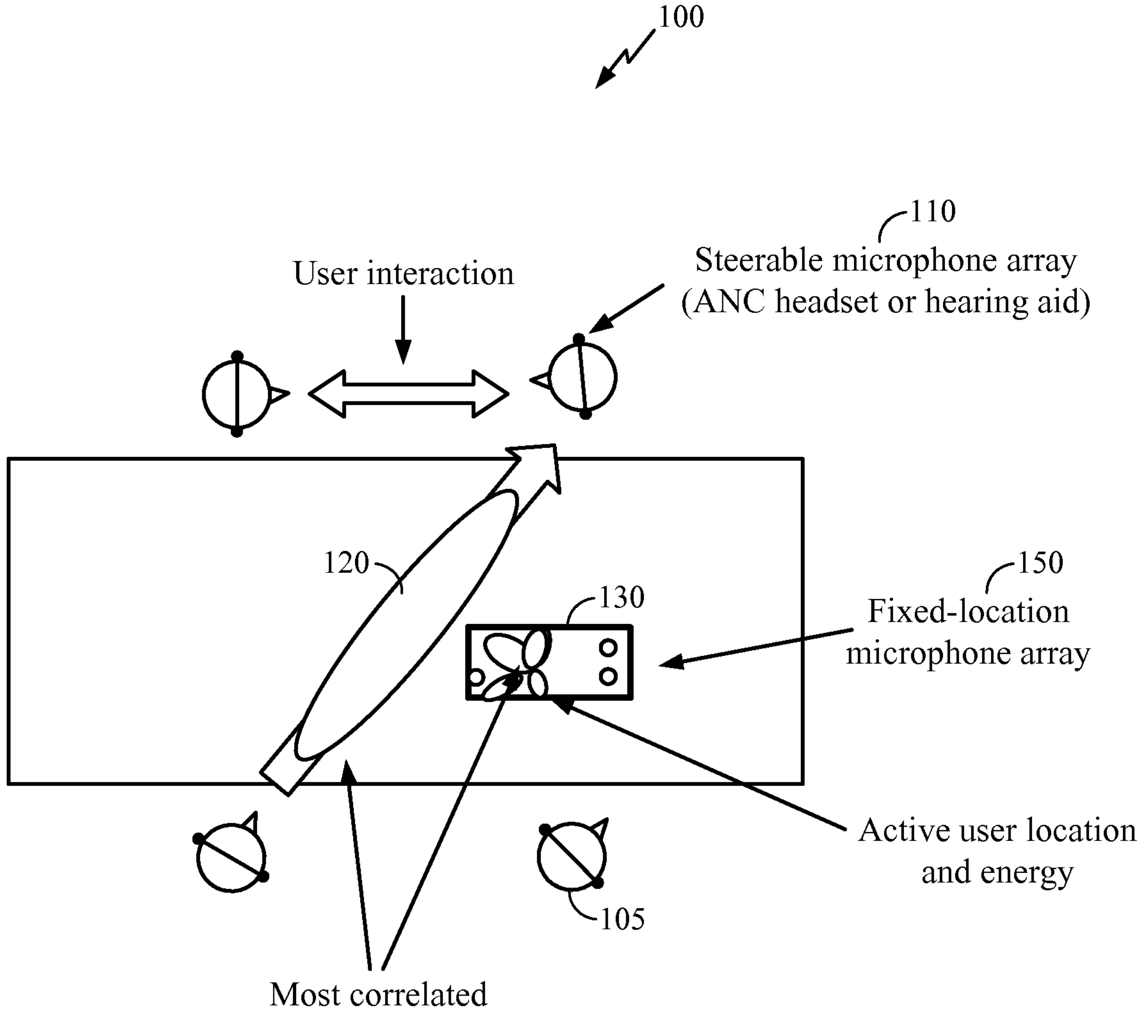

100
110
Steerable microphone array
(ANC headset or hearing aid)
User interaction
120
130
150
Fixed-location
microphone array
Active user location
and energy
105
Most correlated

FIG. 1

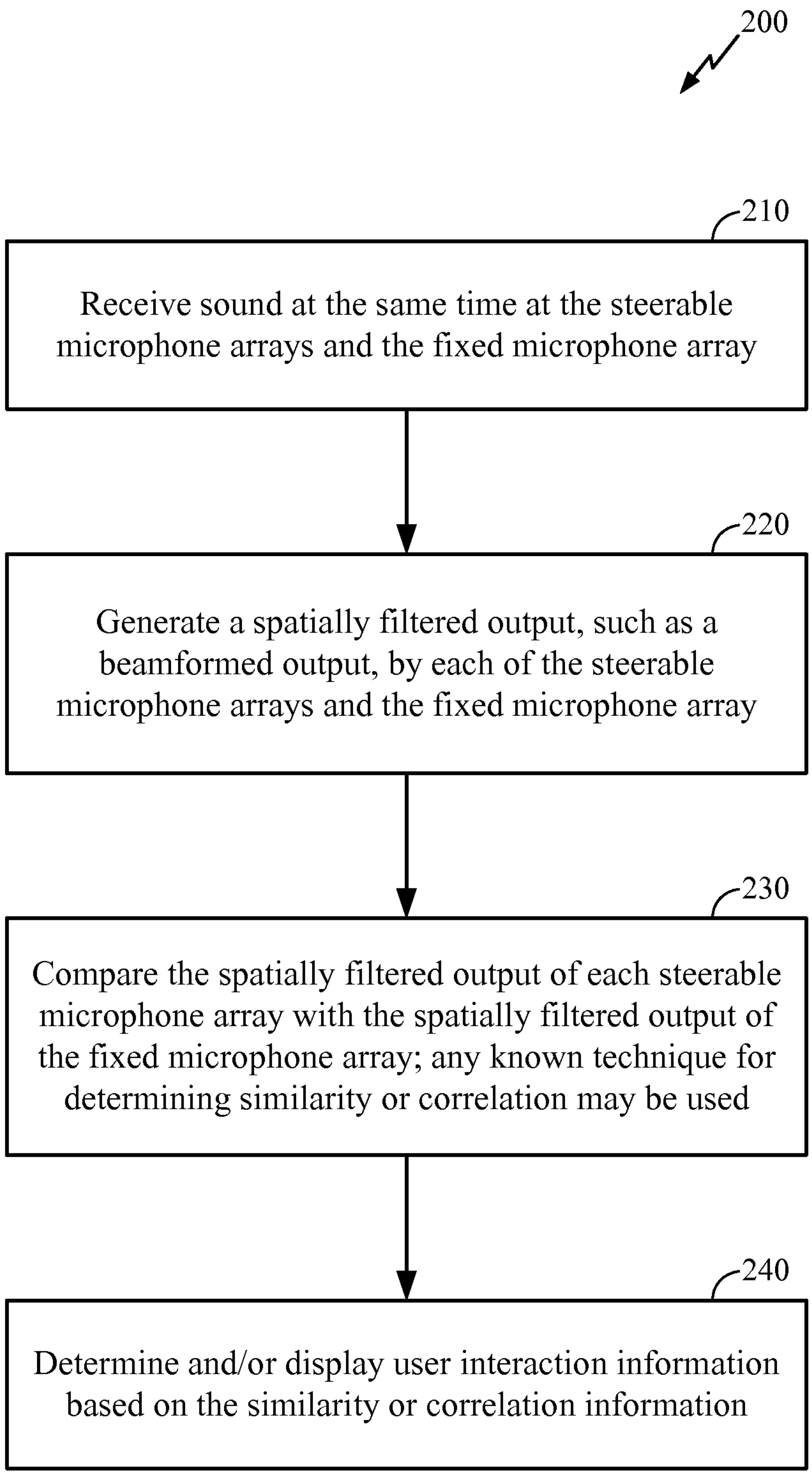

200
210
Receive sound at the same time at the steerable microphone arrays and the fixed microphone array
220
Generate a spatially filtered output, such as a beamformed output, by each of the steerable microphone arrays and the fixed microphone array
230
Compare the spatially filtered output of each steerable microphone array with the spatially filtered output of the fixed microphone array; any known technique for determining similarity or correlation may be used
240
Determine and/or display user interaction information based on the similarity or correlation information

FIG. 2

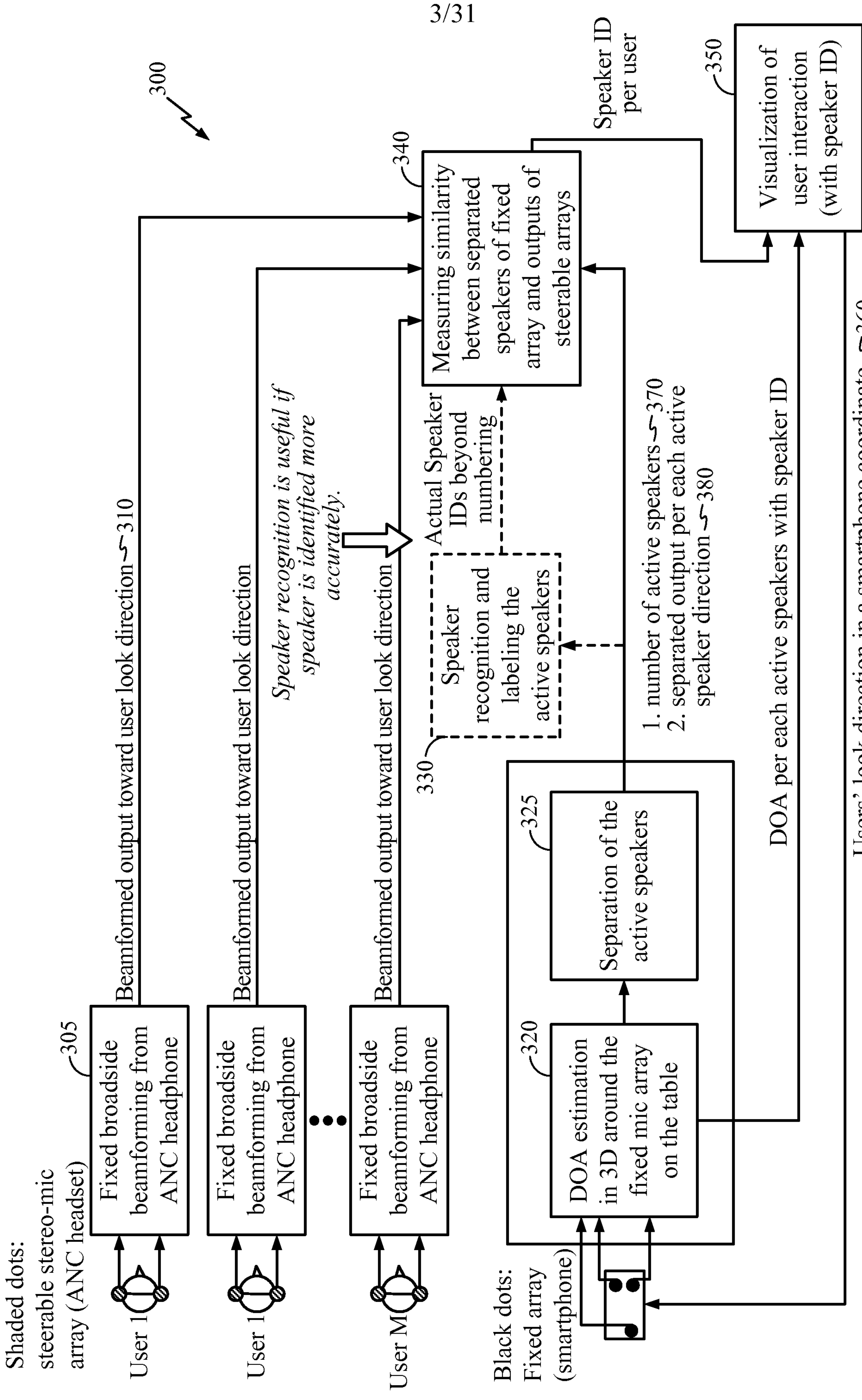

300
Shaded dots: steerable stereo-mic array (ANC headset)
User 1
User 1
User M
305
Fixed broadside beamforming from ANC headphone
Fixed broadside beamforming from ANC headphone
Fixed broadside beamforming from ANC headphone
Beamformed output toward user look direction 310
Beamformed output toward user look direction
Beamformed output toward user look direction
Speaker recognition is useful if speaker is identified more accurately.
330
Speaker recognition and labeling the active speakers
Actual Speaker IDs beyond numbering
340
Measuring similarity between separated speakers of fixed array and outputs of steerable arrays
Speaker ID per user
350
Visualization of user interaction (with speaker ID)
Black dots: Fixed array (smartphone)
320
DOA estimation in 3D around the fixed mic array on the table
325
Separation of the active speakers
1. number of active speakers 370
2. separated output per each active speaker direction 380
DOA per each active speakers with speaker ID
Users' look direction in a smartphone coordinate 360

FIG. 3

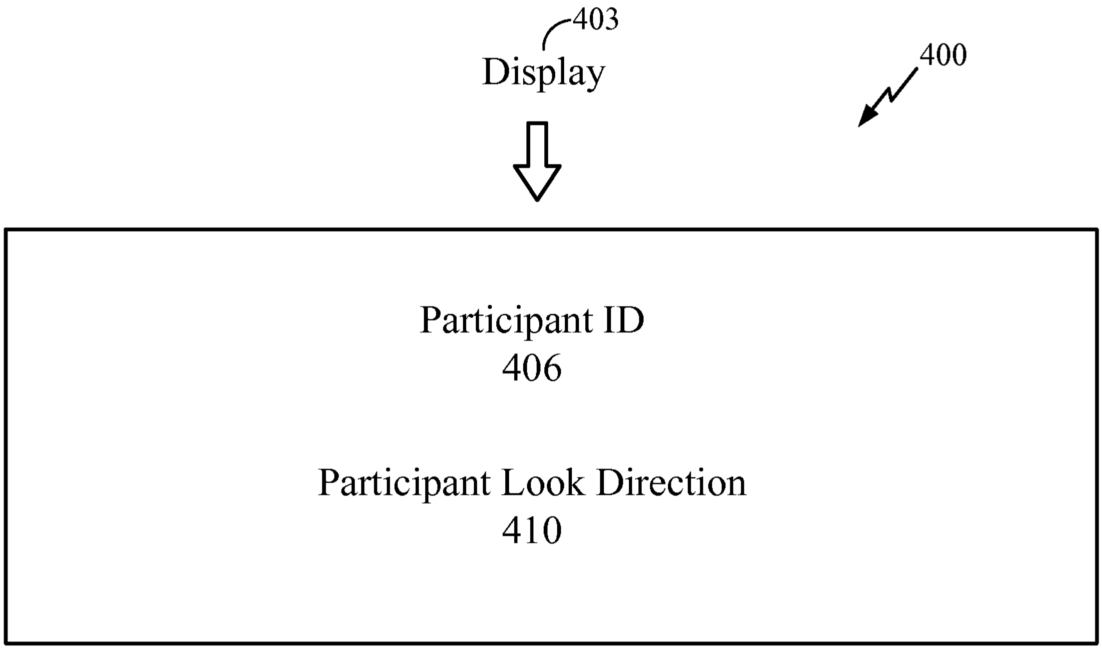

403
Display
400
Participant ID
406
Participant Look Direction
410

FIG. 4

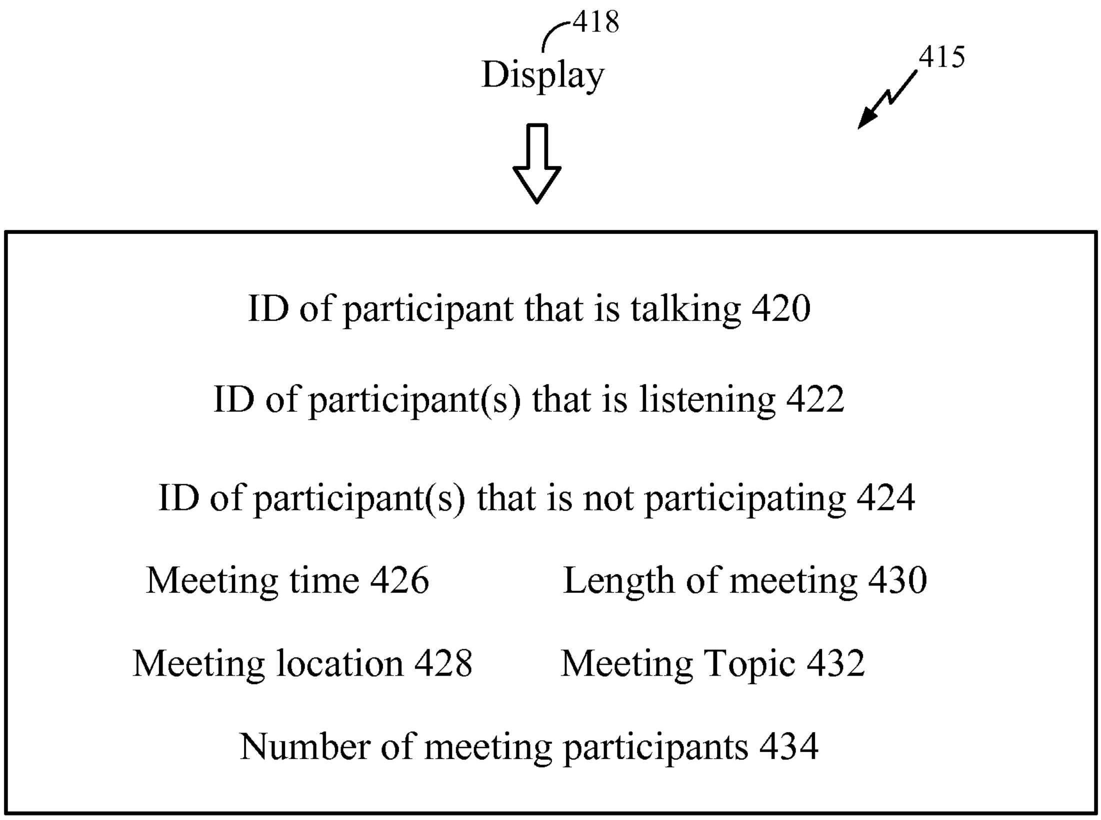

418
Display
415
ID of participant that is talking 420
ID of participant(s) that is listening 422
ID of participant(s) that is not participating 424
Meeting time 426
Length of meeting 430
Meeting location 428
Meeting Topic 432
Number of meeting participants 434

FIG. 5

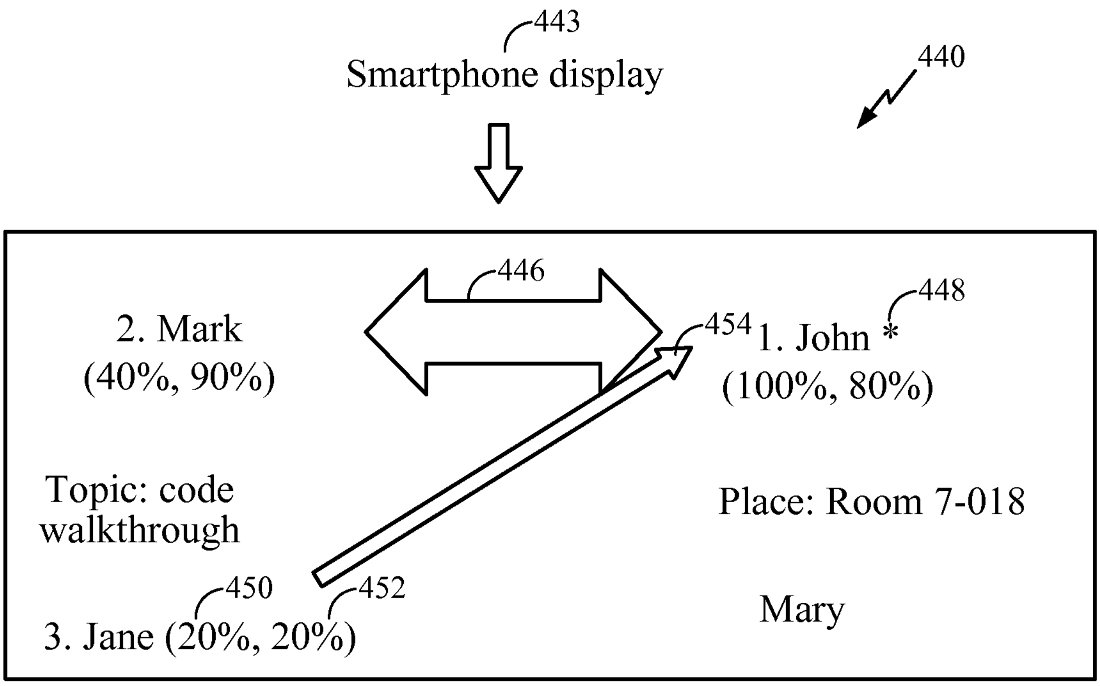

443
Smartphone display
440
446
448
2. Mark
(40%, 90%)
454
1. John *
(100%, 80%)
Topic: code walkthrough
Place: Room 7-018
450
452
3. Jane (20%, 20%)
Mary

FIG. 6

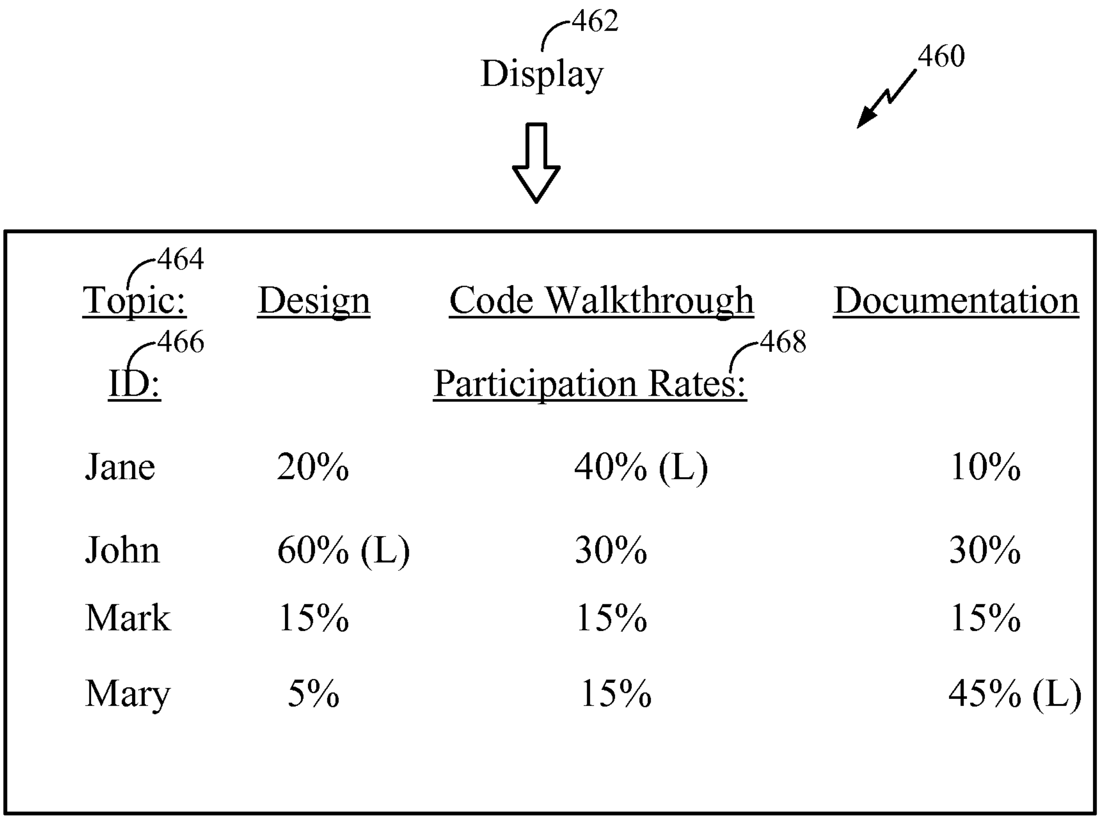

462
Display
460
464
Topic:
Design
Code Walkthrough
Documentation
466
ID:
468
Participation Rates:
Jane 20% 40% (L) 10%
John 60% (L) 30% 30%
Mark 15% 15% 15%
Mary 5% 15% 45% (L)

FIG. 7

472

Display

470

Date: May 3, 2012

474

Topic: Code Walkthrough   User ID: Jane

476

Time →

478

User was looking at:	Mary	Mary	John	John	John	Mark	John	John
User was speaking?	No	No	Yes	Yes	Yes	No	Yes	Yes
% of participants looking at user?	0	0	33	100	100	0	33	66

FIG. 8

482

Display

480

Time Period: May 2011 – November 2011

488

User: Jane   Interaction scale 0 to 10

484 486

ID:	May	June	July	Aug	Sept	Oct	Nov
John	5	5	5	5	8	8	6
Mark	0	0	7	7	8	8	9
Mary	8	8	8	6	6	4	1

FIG. 9

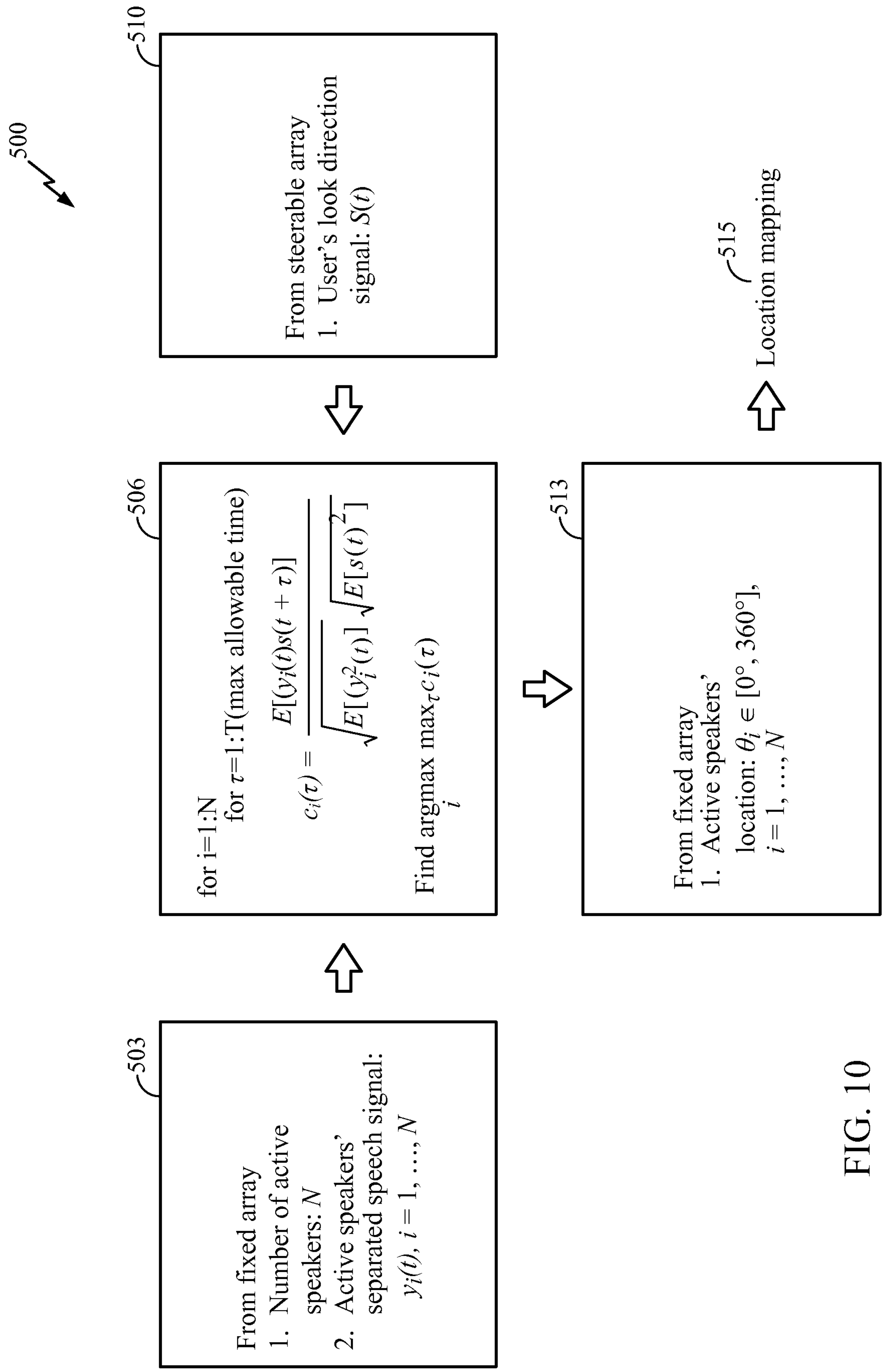

500
510
From steerable array
1. User's look direction signal: $S(t)$
506
for i=1:N
for τ=1:T(max allowable time)
$c_i(\tau) = \frac{E[(y_i(t)s(t+\tau)]}{\sqrt{E[(y_i^2(t)]}\sqrt{E[s(t)^2]}}$
Find $\underset{i}{\text{argmax}}\ \max_\tau c_i(\tau)$
503
From fixed array
1. Number of active speakers: N
2. Active speakers' separated speech signal: $y_i(t)$, $i = 1, \ldots, N$
513
From fixed array
1. Active speakers' location: $\theta_i \in [0°, 360°]$, $i = 1, \ldots, N$
515
Location mapping

FIG. 10

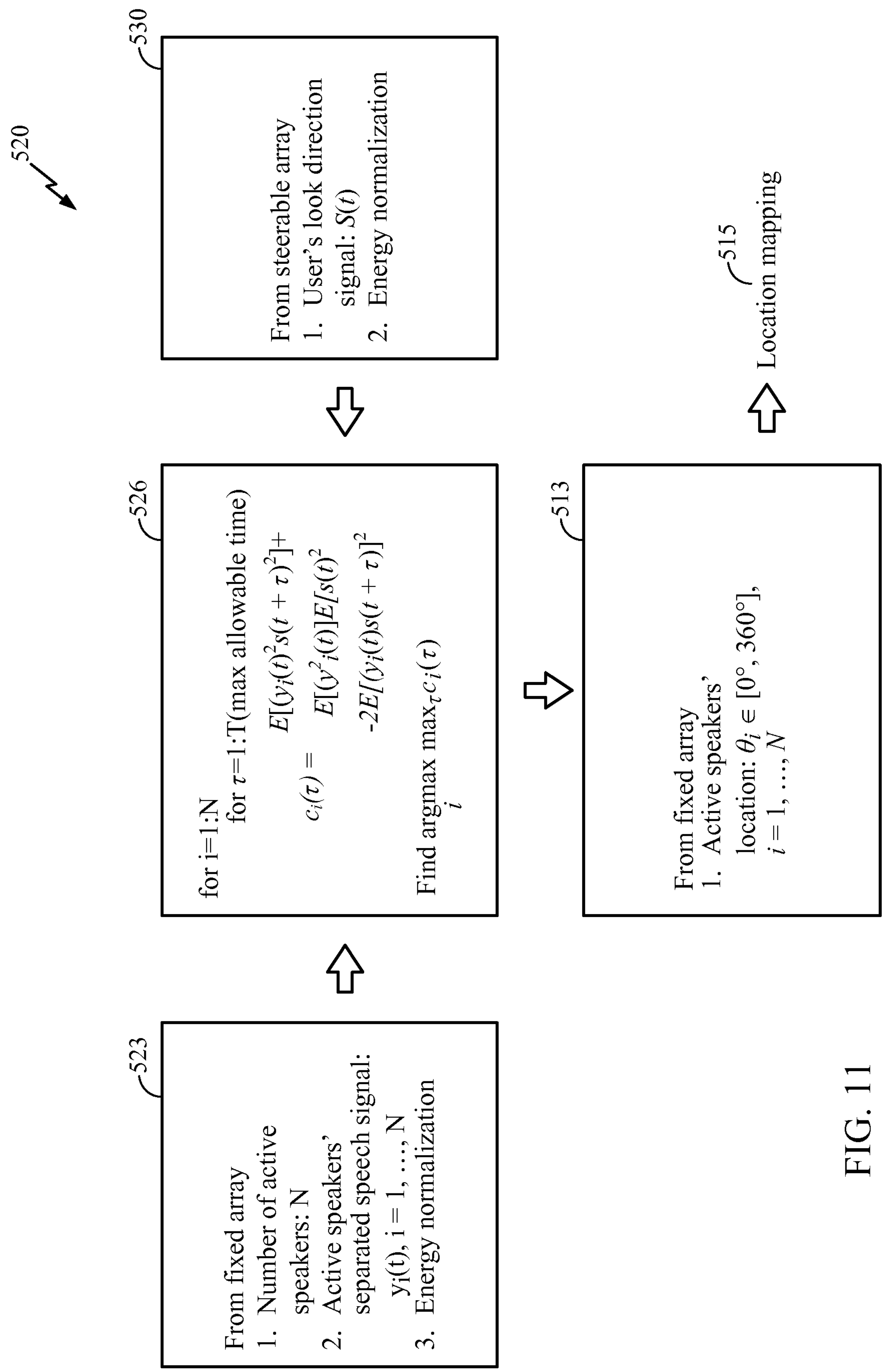

520
523
From fixed array
1. Number of active speakers: N
2. Active speakers' separated speech signal: yi(t), i = 1, ..., N
3. Energy normalization
526
for i=1:N
for τ=1:T(max allowable time)
ci(τ) = E[(yi(t)2s(t + τ)2]+ E[(y2i(t)]E[s(t)2 -2E[(yi(t)s(t + τ)]2
Find argmax maxτ ci(τ) i
530
From steerable array
1. User's look direction signal: S(t)
2. Energy normalization
513
From fixed array
1. Active speakers' location: θi ∈ [0°, 360°], i = 1, ..., N
515
Location mapping

FIG. 11

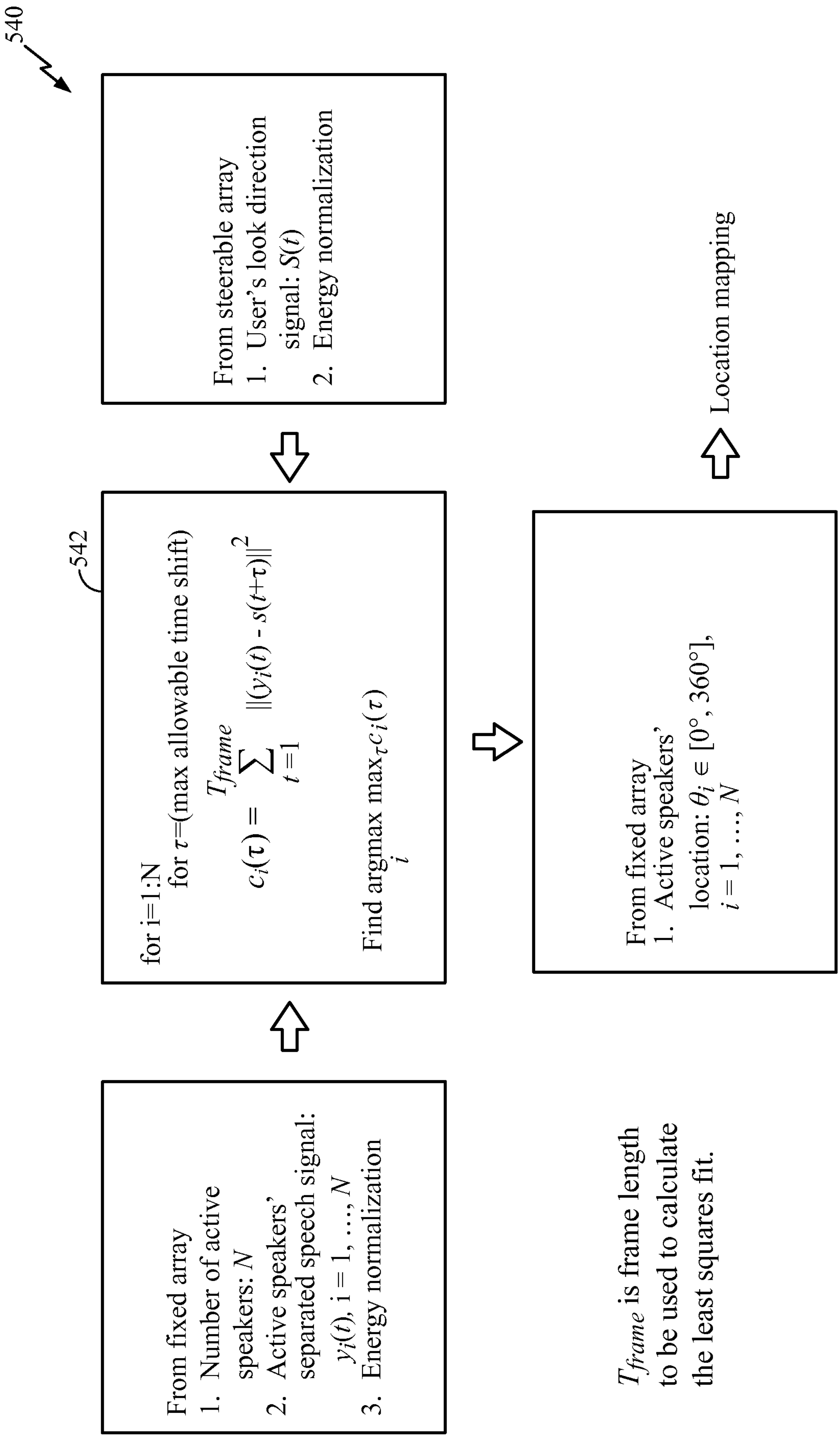

540
From steerable array
1. User's look direction signal: $S(t)$
2. Energy normalization
542
for i=1:N
for τ=(max allowable time shift)
$c_i(\tau) = \sum_{t=1}^{T_{frame}} \|(y_i(t) - s(t+\tau))\|^2$
Find $\underset{i}{\text{argmax}}\ \max_{\tau} c_i(\tau)$
From fixed array
1. Number of active speakers: N
2. Active speakers' separated speech signal: $y_i(t)$, i = 1, ..., N
3. Energy normalization
From fixed array
1. Active speakers' location: $\theta_i \in [0°, 360°]$, $i = 1, \ldots, N$
Location mapping
T_{frame} is frame length to be used to calculate the least squares fit.

FIG. 12

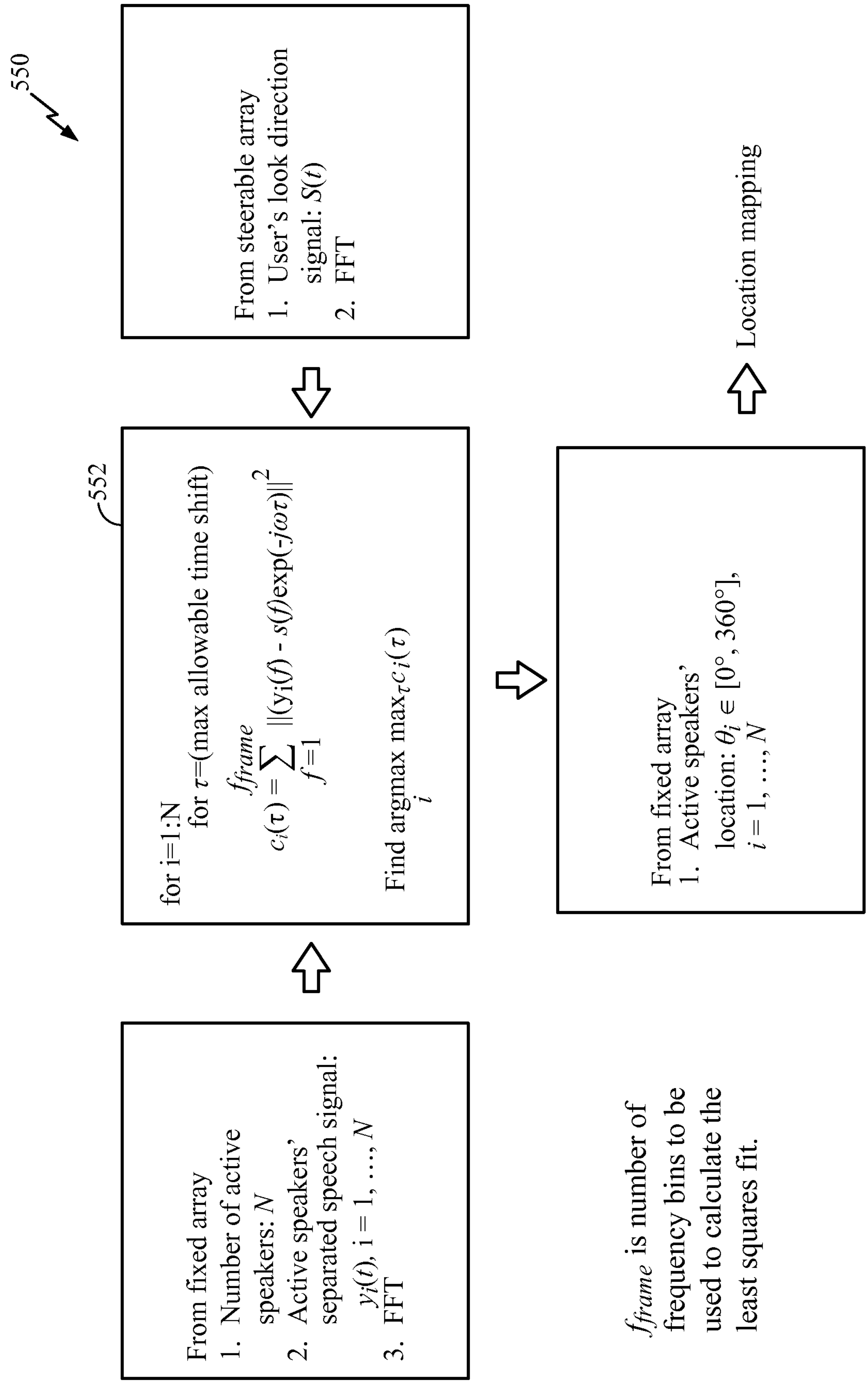

550
From steerable array
1. User's look direction signal: $S(t)$
2. FFT
552
for i=1:N
for τ=(max allowable time shift)
$c_i(\tau) = \sum_{f=1}^{f_{frame}} \|(y_i(f) - s(f)\exp(-j\omega\tau)\|^2$
Find $\arg\max_i \max_\tau c_i(\tau)$
From fixed array
1. Number of active speakers: N
2. Active speakers' separated speech signal: $y_i(t)$, i = 1, …, N
3. FFT
From fixed array
1. Active speakers' location: $\theta_i \in [0°, 360°]$, $i = 1, …, N$
Location mapping
f_{frame} is number of frequency bins to be used to calculate the least squares fit.

FIG. 13

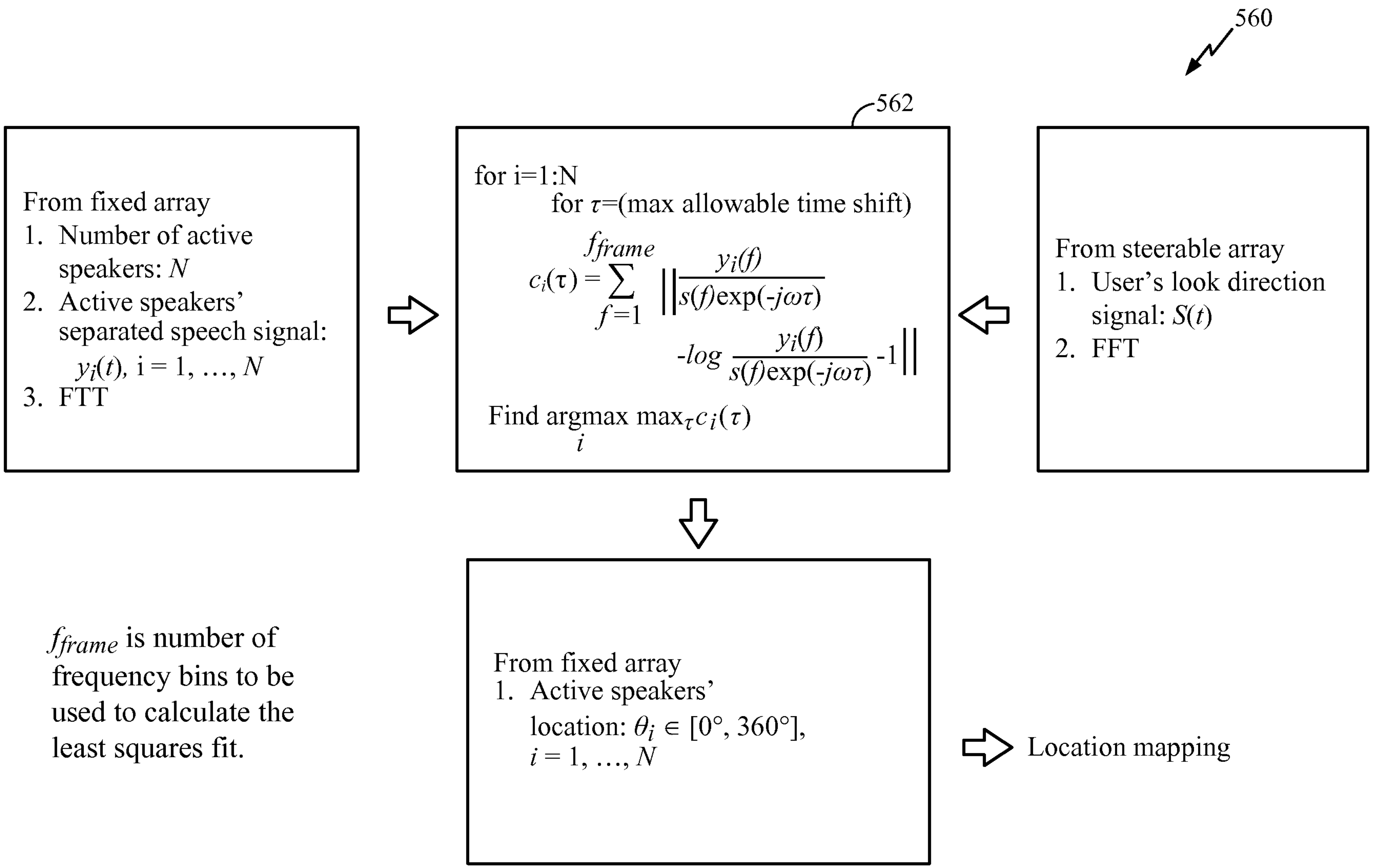

560
562
From fixed array
1. Number of active speakers: N
2. Active speakers' separated speech signal: $y_i(t)$, i = 1, …, N
3. FTT
for i=1:N
for τ=(max allowable time shift)
$c_i(\tau) = \sum_{f=1}^{f_{frame}} \left\| \frac{y_i(f)}{s(f)\exp(-j\omega\tau)} - \log \frac{y_i(f)}{s(f)\exp(-j\omega\tau)} - 1 \right\|$
Find $\underset{i}{\operatorname{argmax}} \max_\tau c_i(\tau)$
From steerable array
1. User's look direction signal: $S(t)$
2. FFT
From fixed array
1. Active speakers' location: $\theta_i \in [0°, 360°]$, $i = 1, \ldots, N$
Location mapping
f_{frame} is number of frequency bins to be used to calculate the least squares fit.

FIG. 14

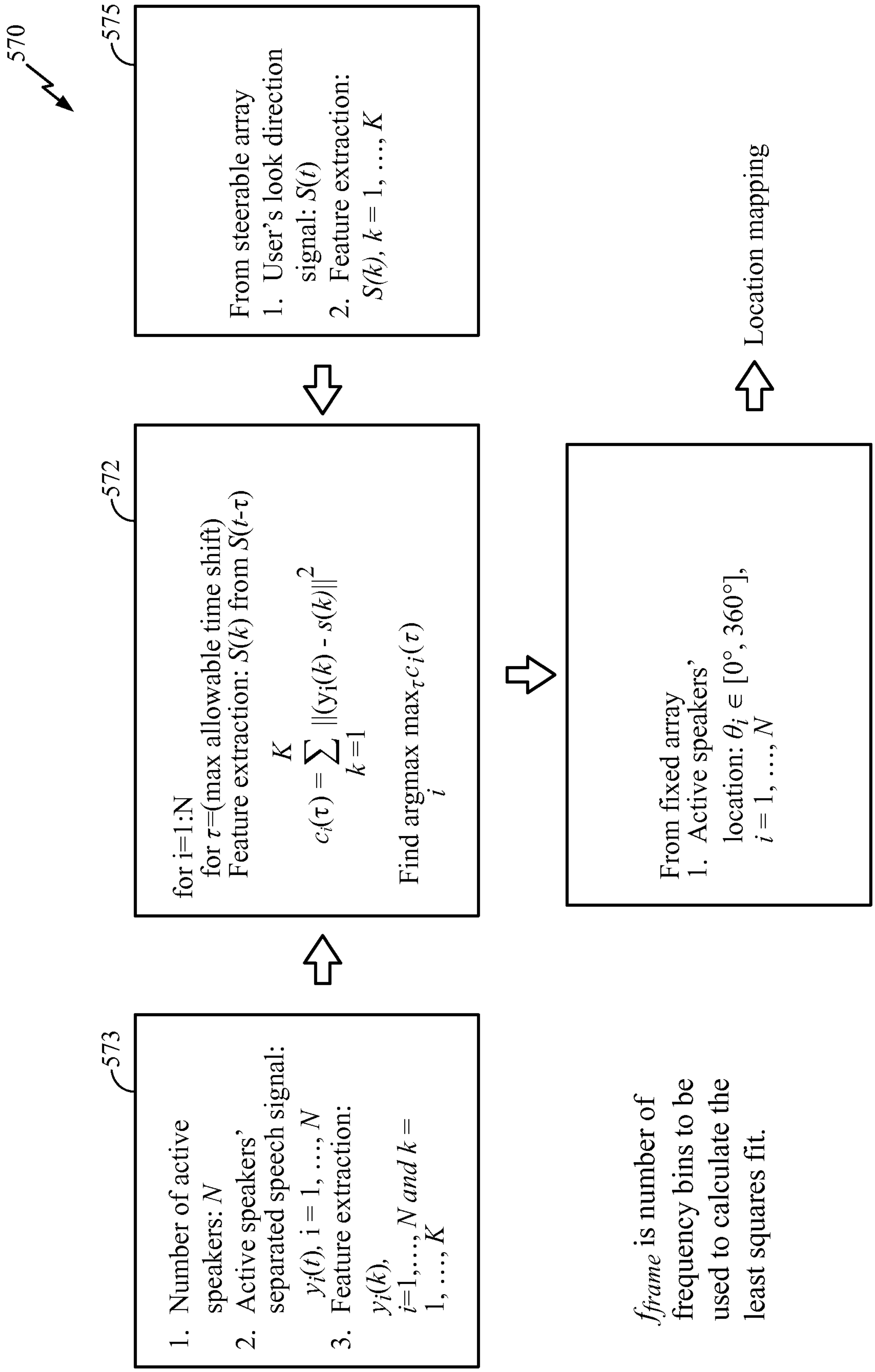

570
573
1. Number of active speakers: N
2. Active speakers' separated speech signal: $y_i(t)$, i = 1, ..., N
3. Feature extraction: $y_i(k)$, i=1,...,N and k = 1, ..., K
572
for i=1:N
for τ=(max allowable time shift)
Feature extraction: S(k) from S(t-τ)
$c_i(\tau) = \sum_{k=1}^{K} \|(y_i(k) - s(k))\|^2$
Find $\arg\max_i \max_\tau c_i(\tau)$
575
From steerable array
1. User's look direction signal: S(t)
2. Feature extraction: S(k), k = 1, ..., K
From fixed array
1. Active speakers' location: $\theta_i \in [0°, 360°]$, i = 1, ..., N
Location mapping
f_{frame} is number of frequency bins to be used to calculate the least squares fit.

FIG. 15

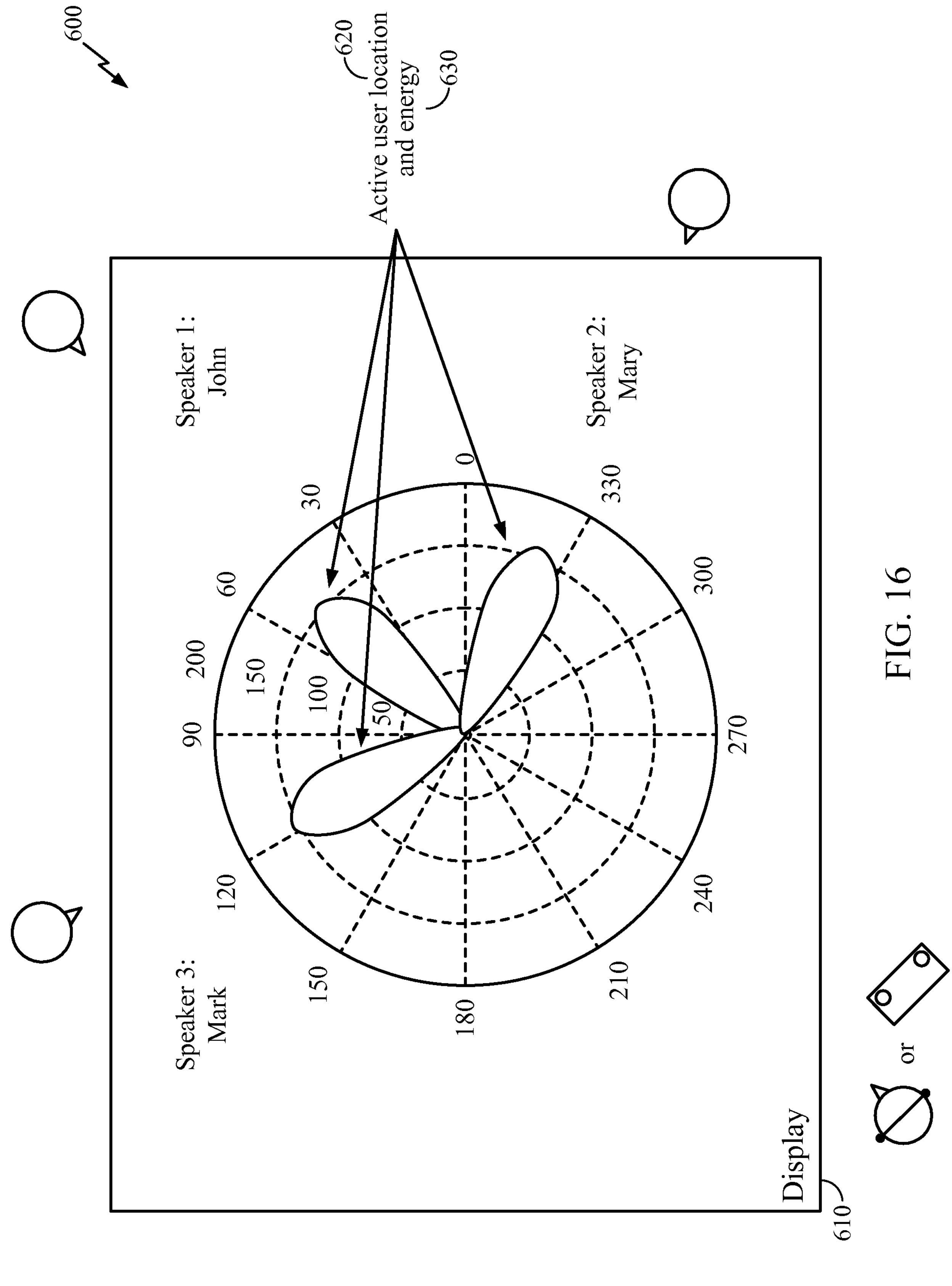

600
620
Active user location and energy
630
Speaker 1: John
Speaker 2: Mary
Speaker 3: Mark
0
30
60
90
120
150
180
210
240
270
300
330
50
100
150
200
Display
610
or

FIG. 16

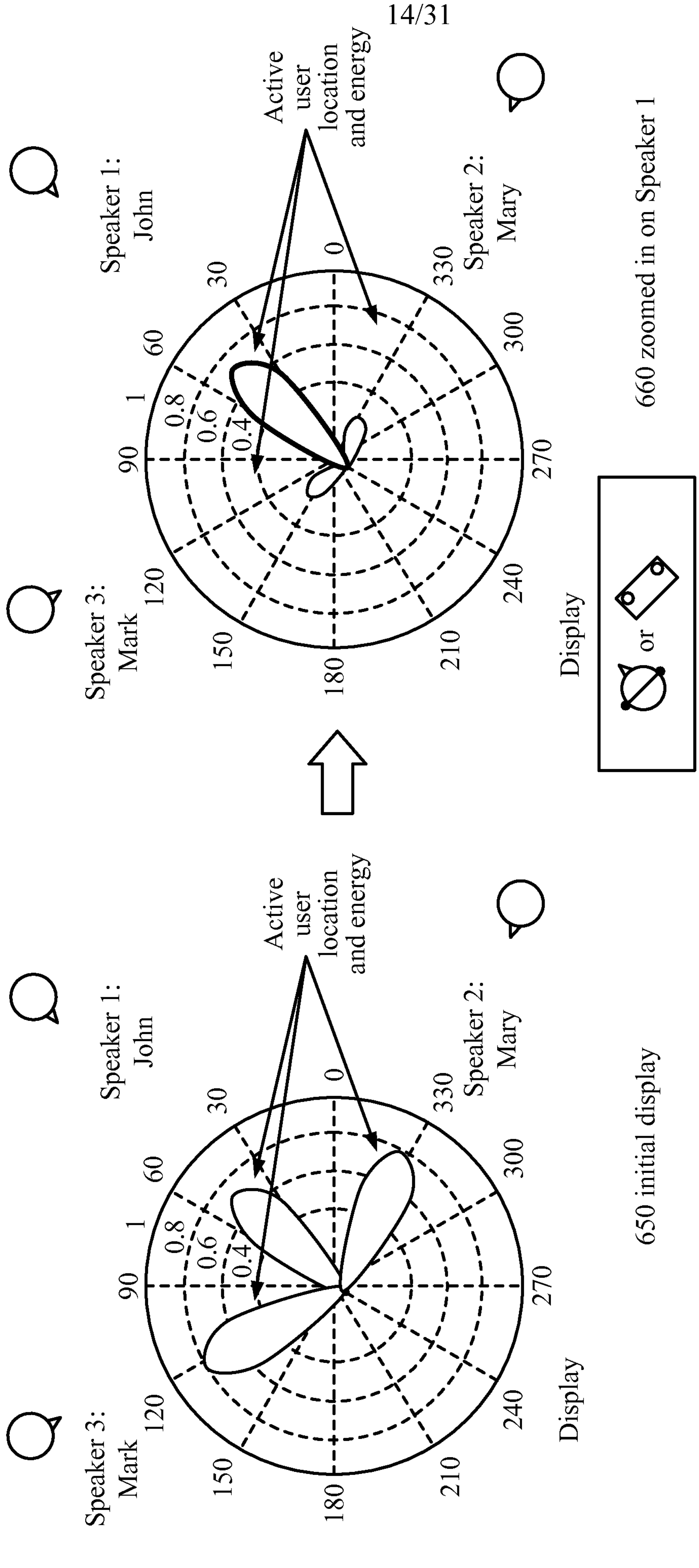

Active user location and energy
Speaker 1: John
Speaker 2: Mary
Speaker 3: Mark
0
30
60
90
120
150
180
210
240
270
300
330
1
0.8
0.6
0.4
Display
or
660 zoomed in on Speaker 1
Active user location and energy
Speaker 1: John
Speaker 2: Mary
Speaker 3: Mark
Display
650 initial display

FIG. 17

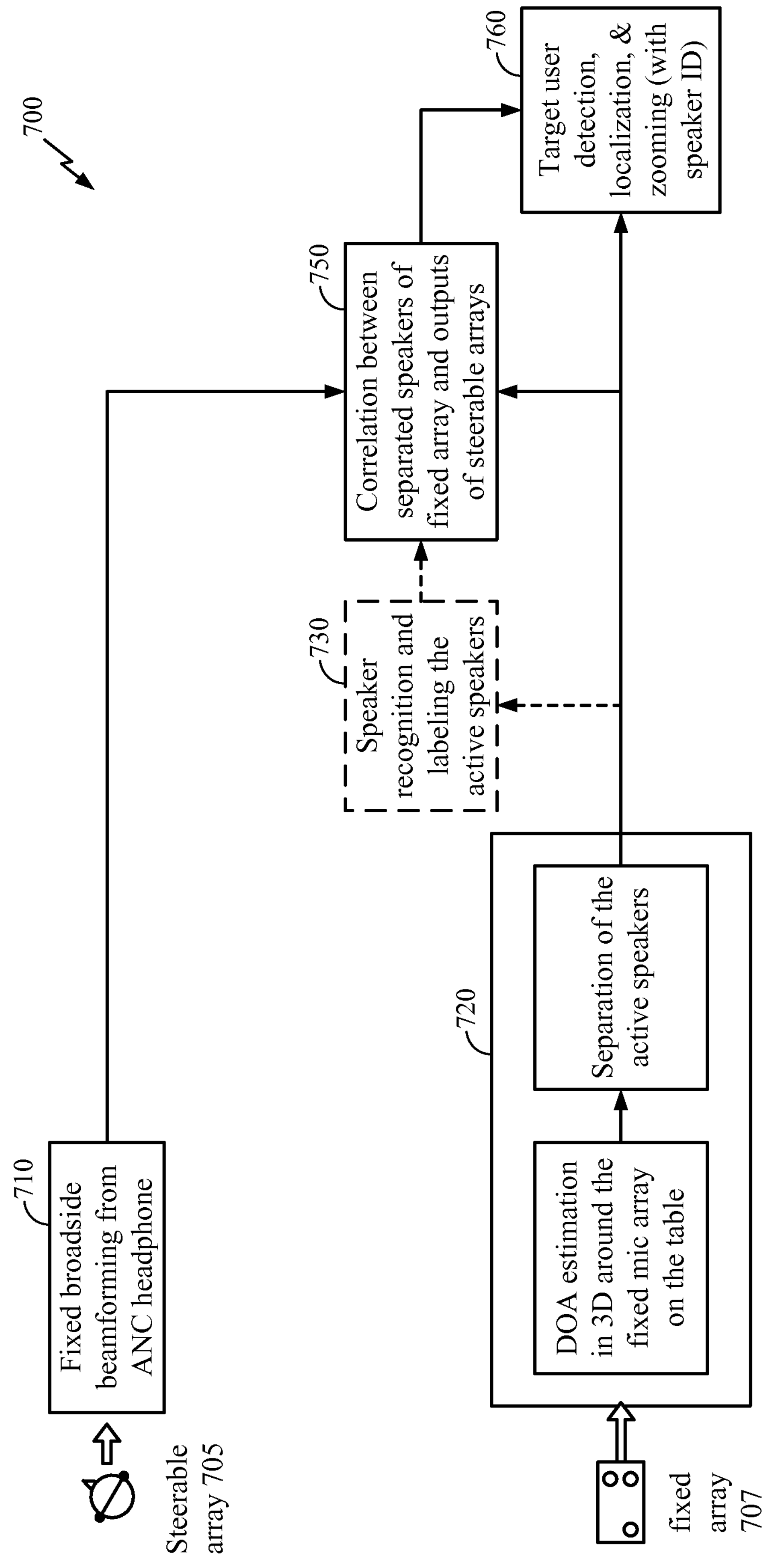

700
Steerable array 705
710
Fixed broadside beamforming from ANC headphone
fixed array 707
720
DOA estimation in 3D around the fixed mic array on the table
Separation of the active speakers
730
Speaker recognition and labeling the active speakers
750
Correlation between separated speakers of fixed array and outputs of steerable arrays
760
Target user detection, localization, & zooming (with speaker ID)

FIG. 18

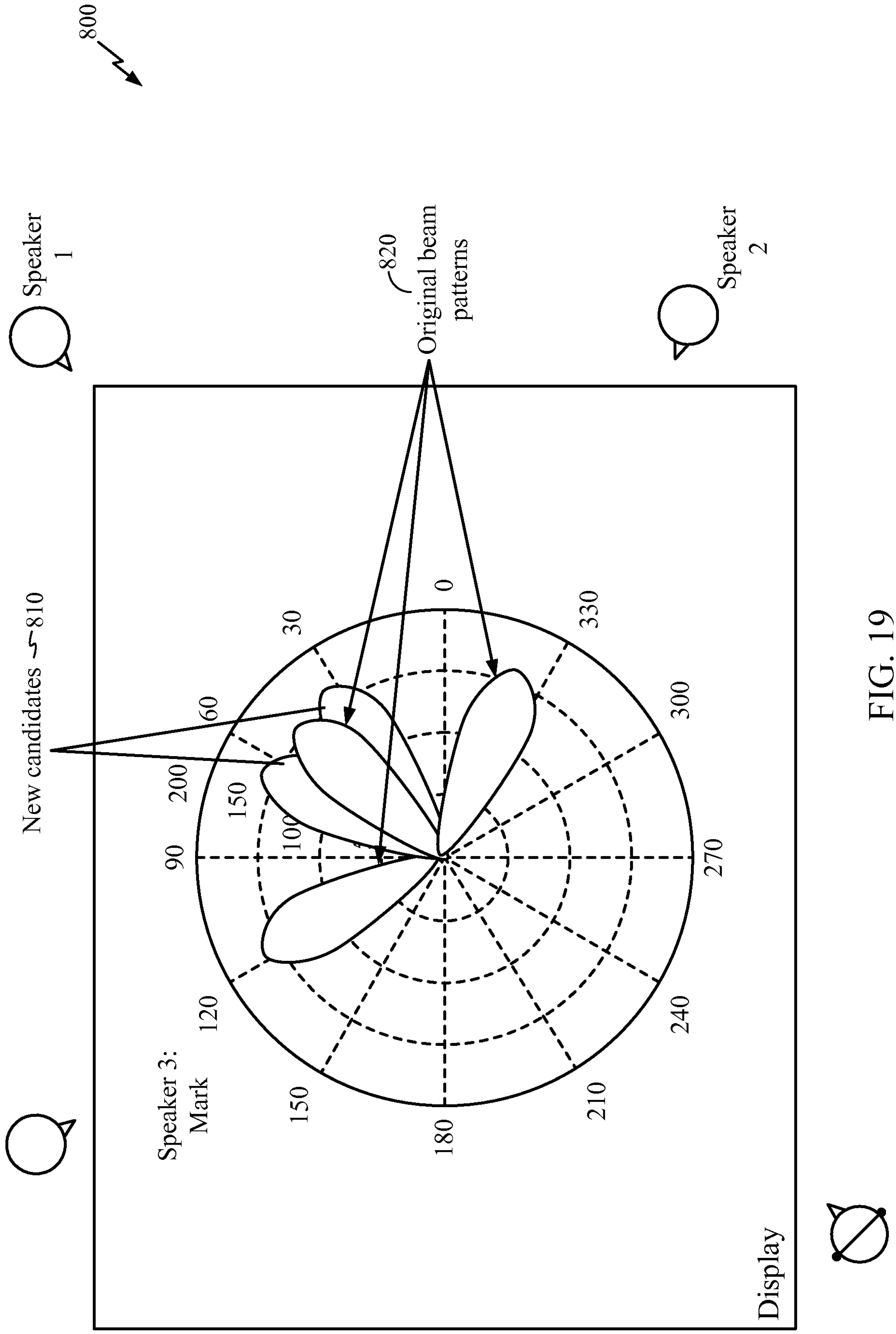

800
Speaker 1
Speaker 2
Original beam patterns
820
New candidates
810
Speaker 3: Mark
Display
0
30
60
90
120
150
180
210
240
270
300
330
100
150
200

FIG. 19

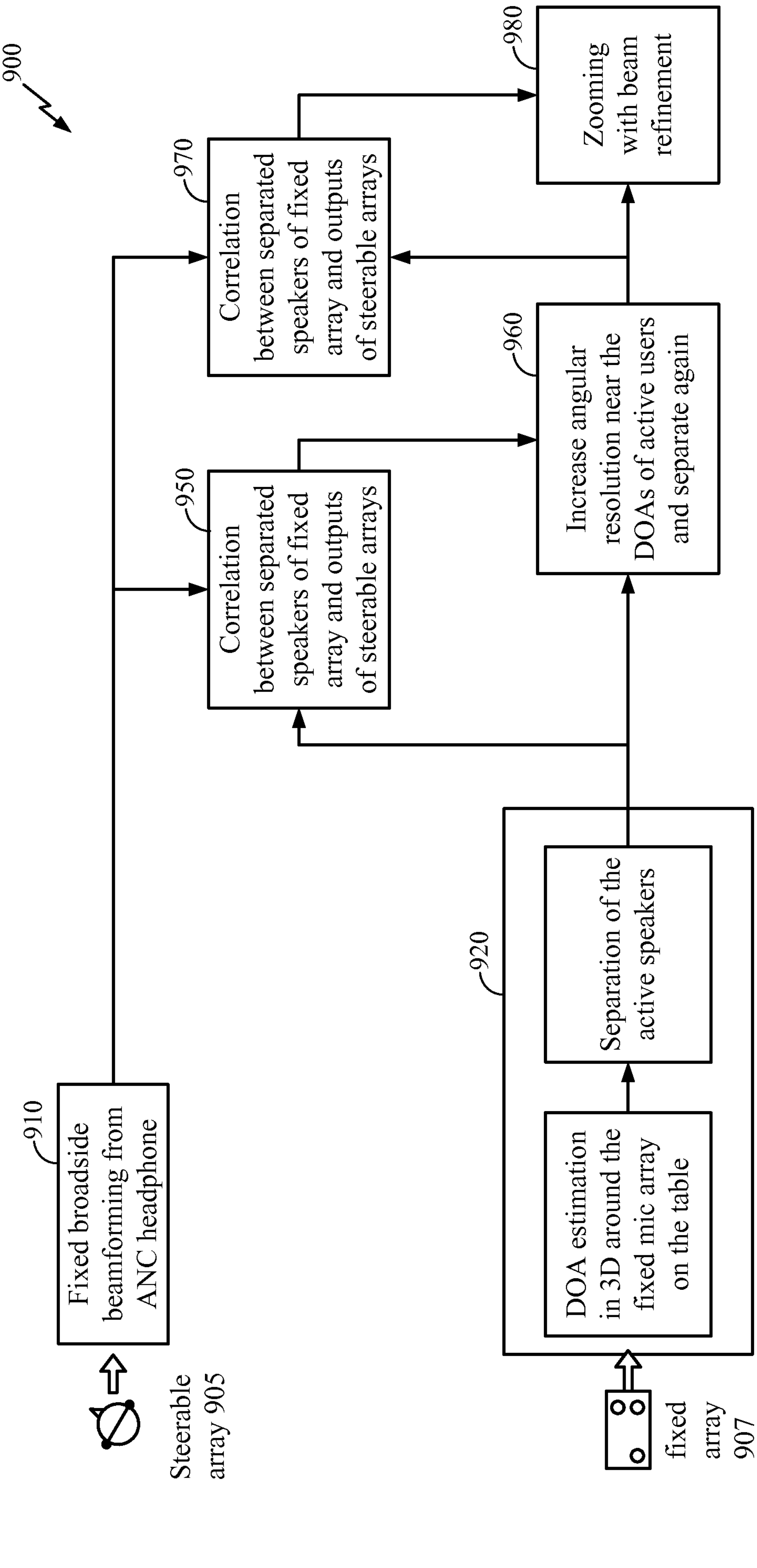

900
910
Fixed broadside beamforming from ANC headphone
Steerable array 905
920
DOA estimation in 3D around the fixed mic array on the table
Separation of the active speakers
fixed array 907
950
Correlation between separated speakers of fixed array and outputs of steerable arrays
960
Increase angular resolution near the DOAs of active users and separate again
970
Correlation between separated speakers of fixed array and outputs of steerable arrays
980
Zooming with beam refinement

FIG. 20

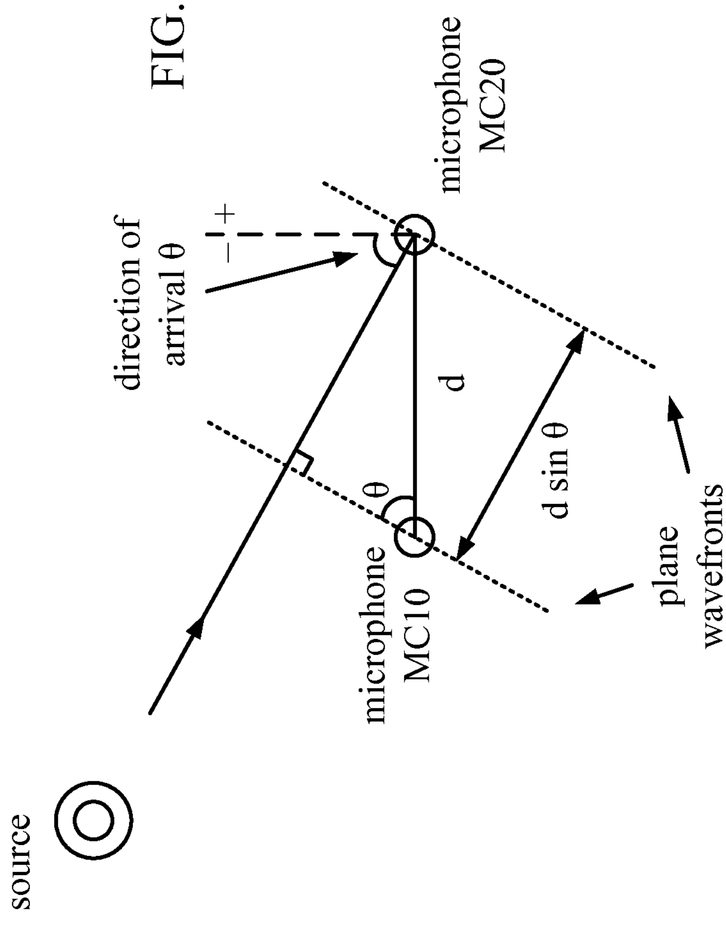

source
microphone
MC10
microphone
MC20
direction of
arrival θ
–
+
d
θ
d sin θ
plane
wavefronts

FIG. 21

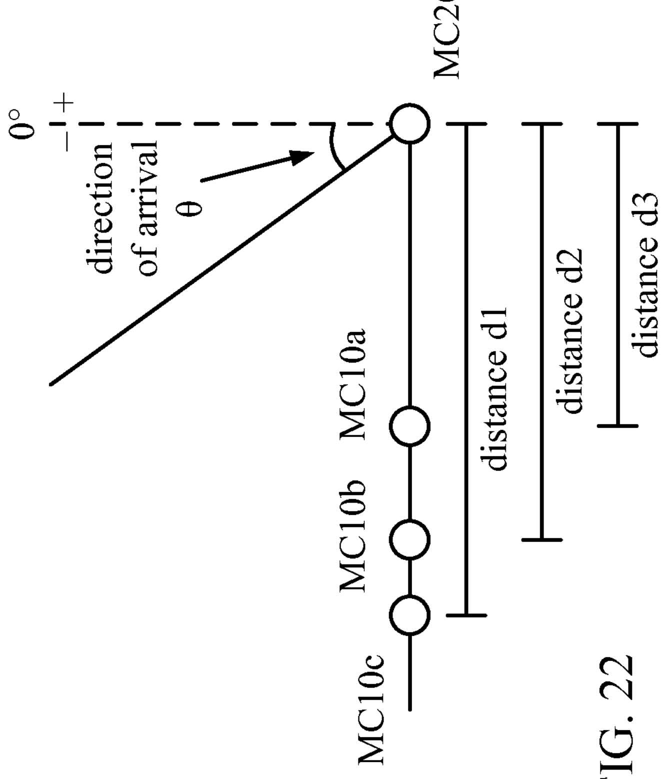

0°
–
+
direction
of arrival
θ
MC20
MC10a
MC10b
MC10c
distance d1
distance d2
distance d3

FIG. 22

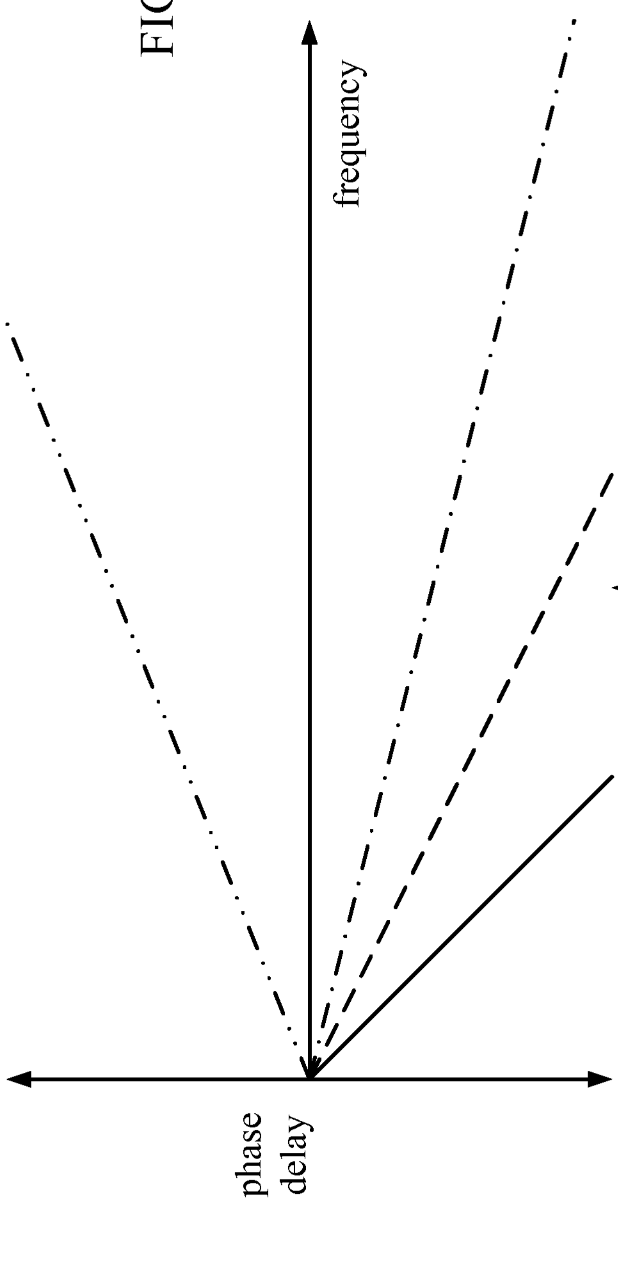


FIG. 23

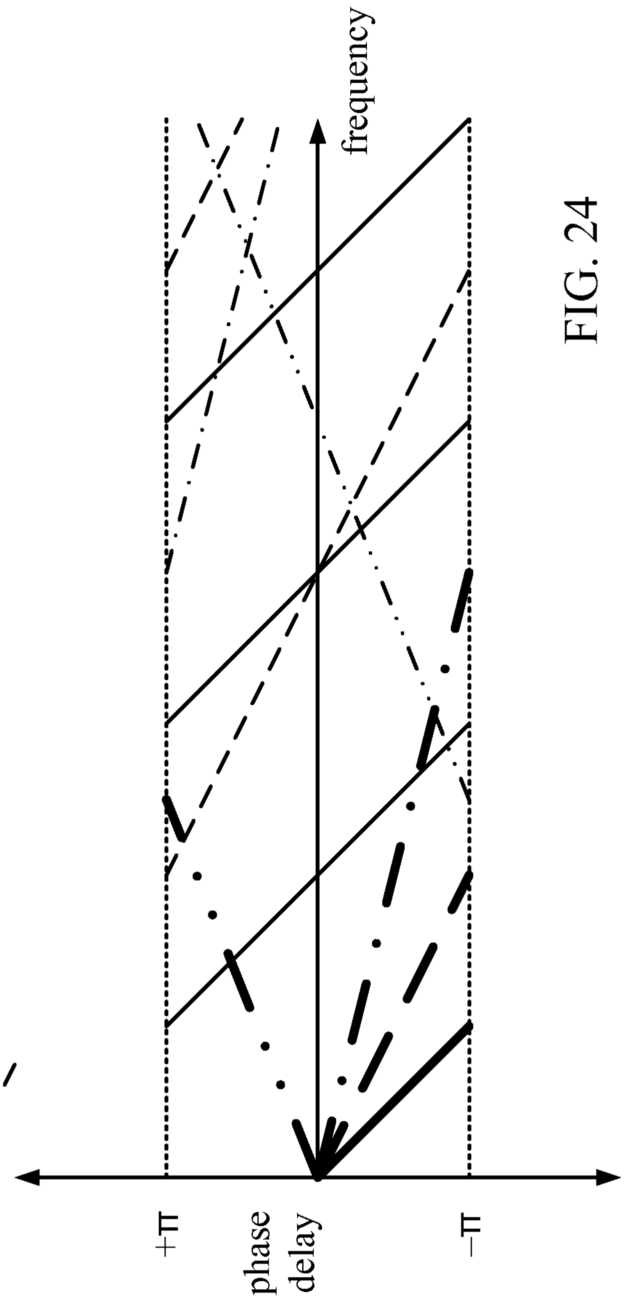


FIG. 24

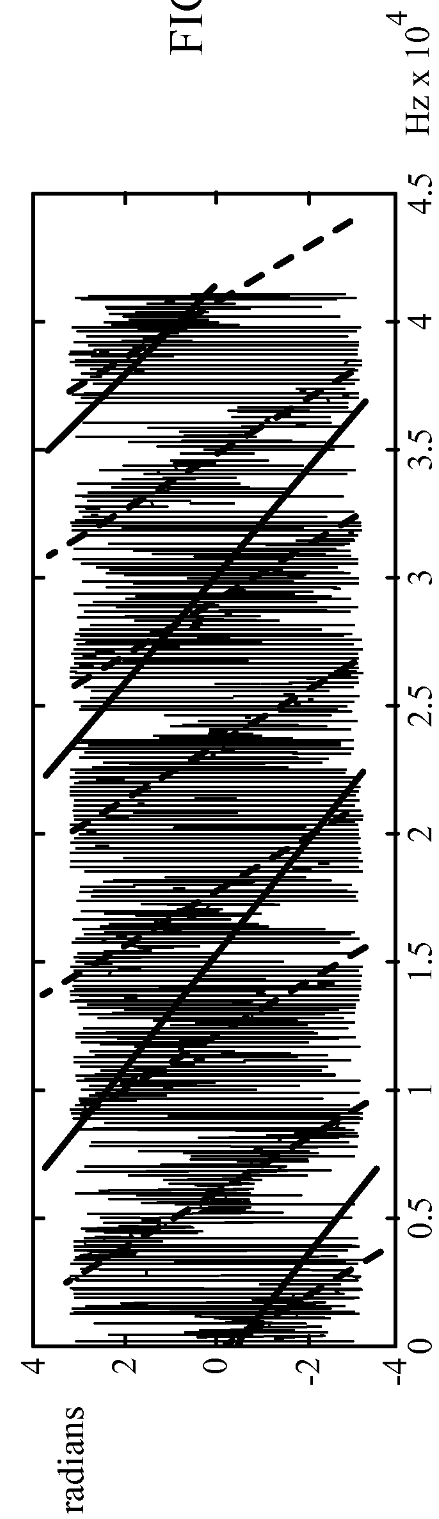


FIG. 25

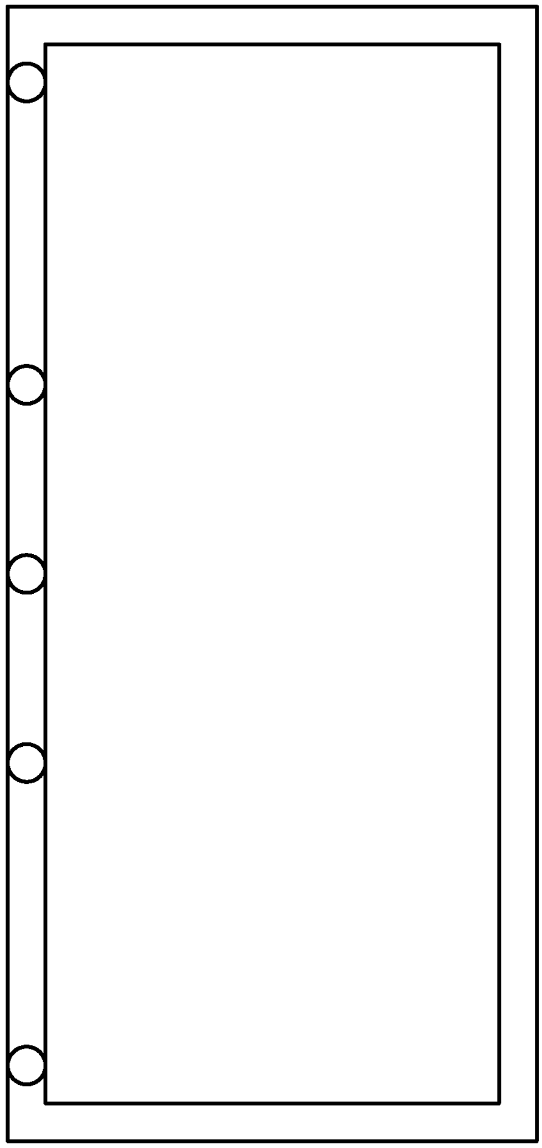


FIG. 26

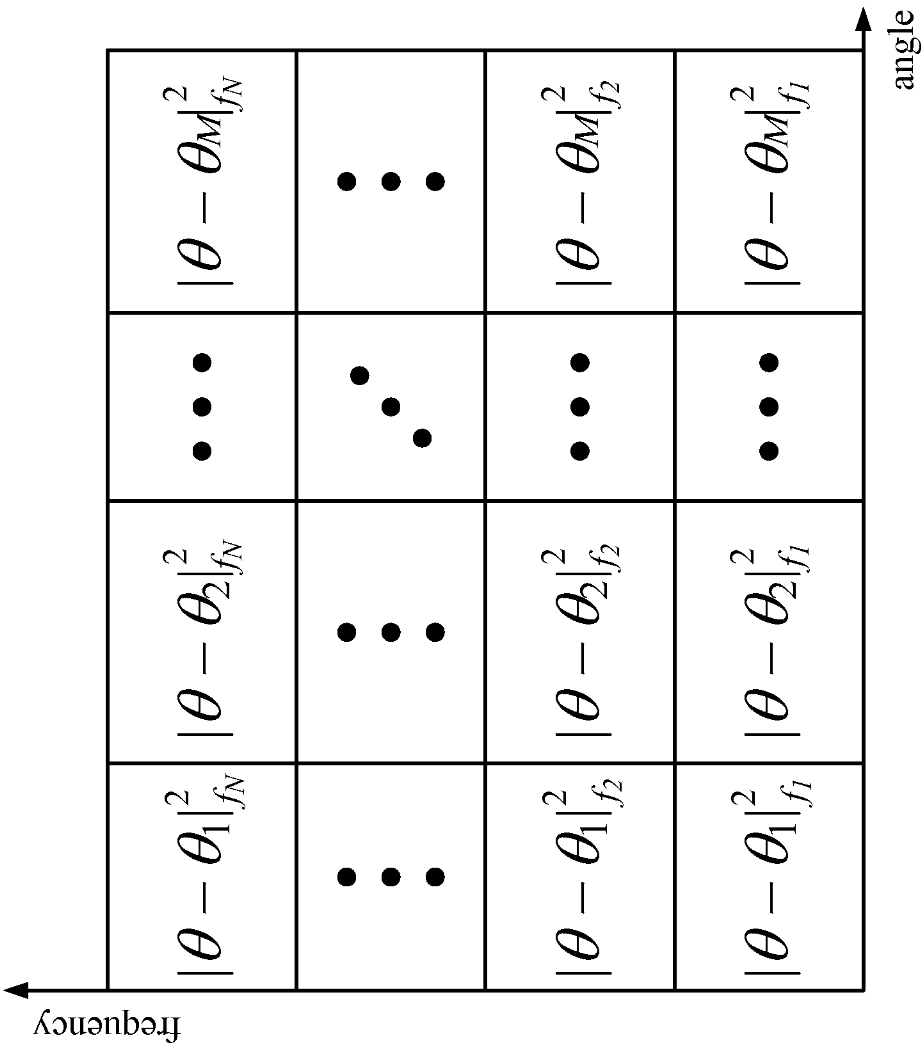


FIG. 27

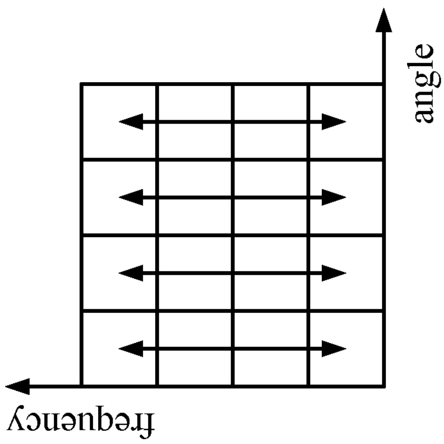


FIG. 28

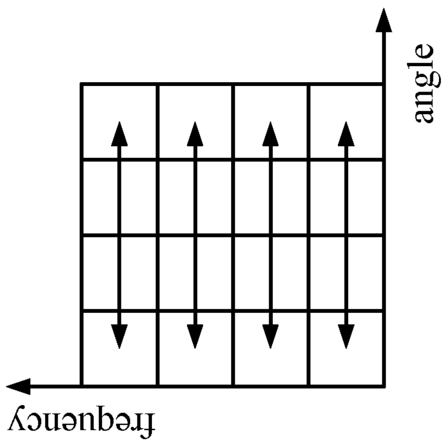


FIG. 29

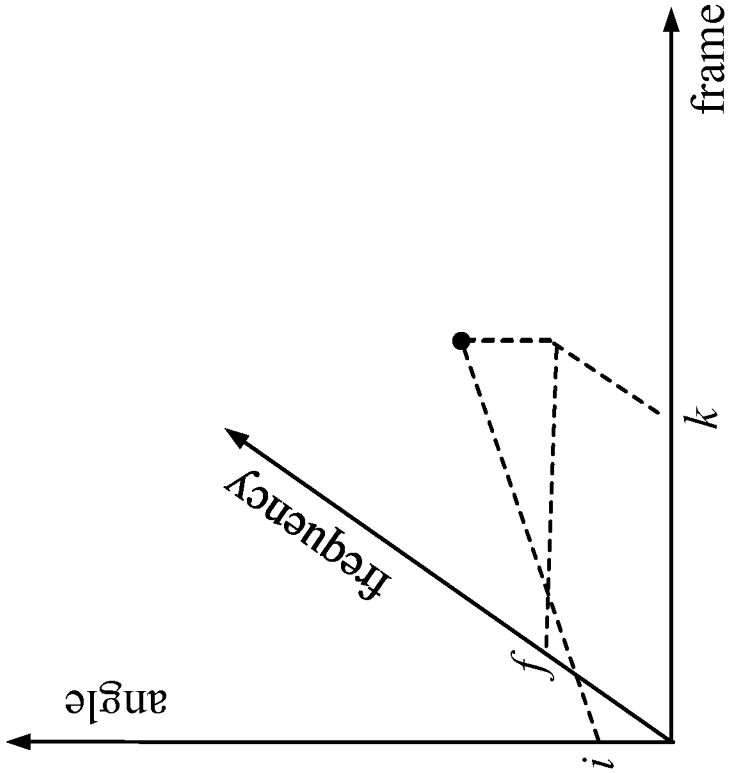


FIG. 31

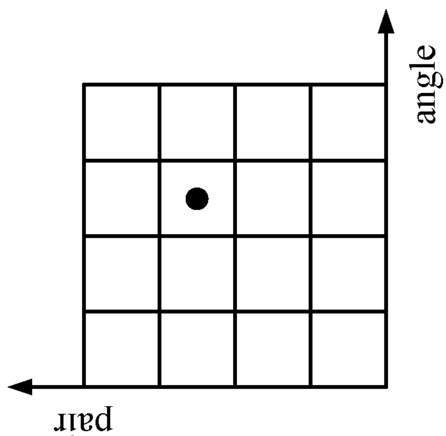


FIG. 30

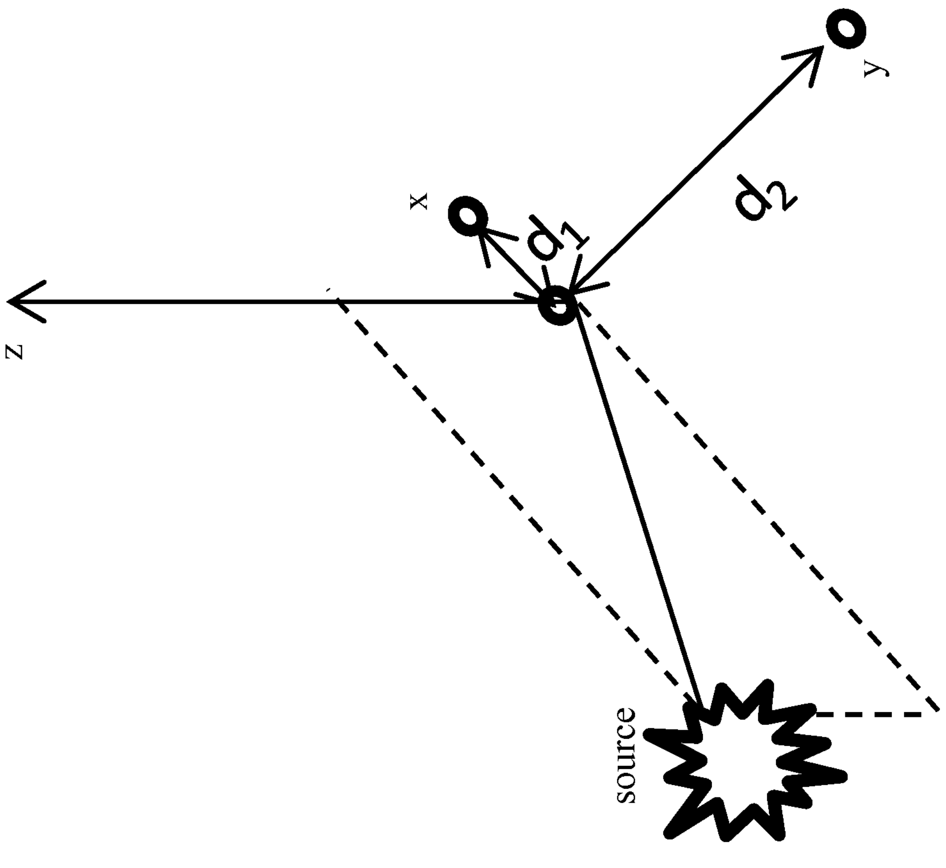

z
x
y
d_1
d_2
source

FIG. 32

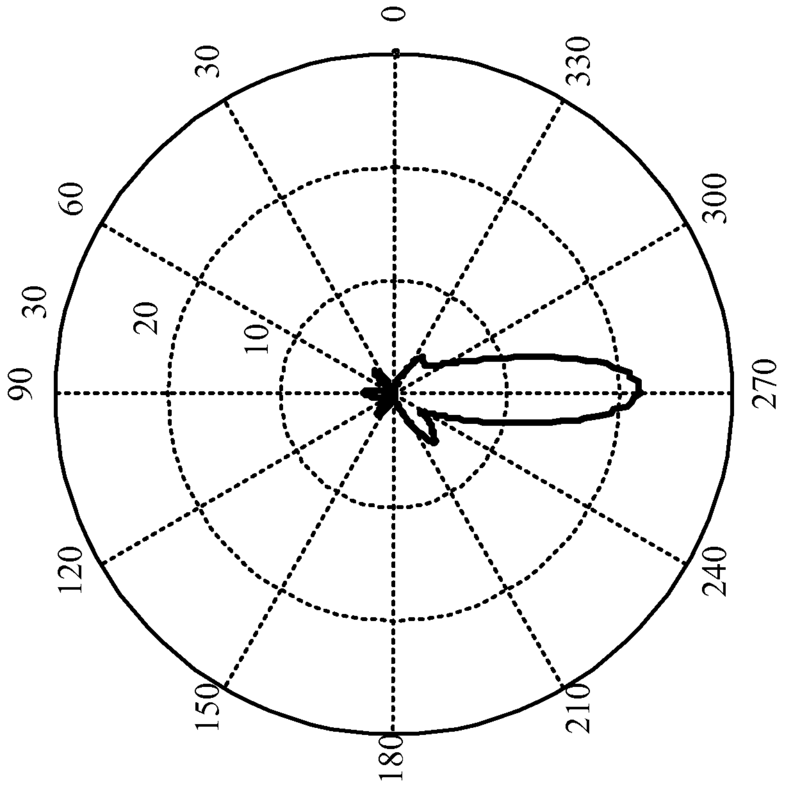

0
30
60
90
120
150
180
210
240
270
300
330
10
20
30

FIG. 33

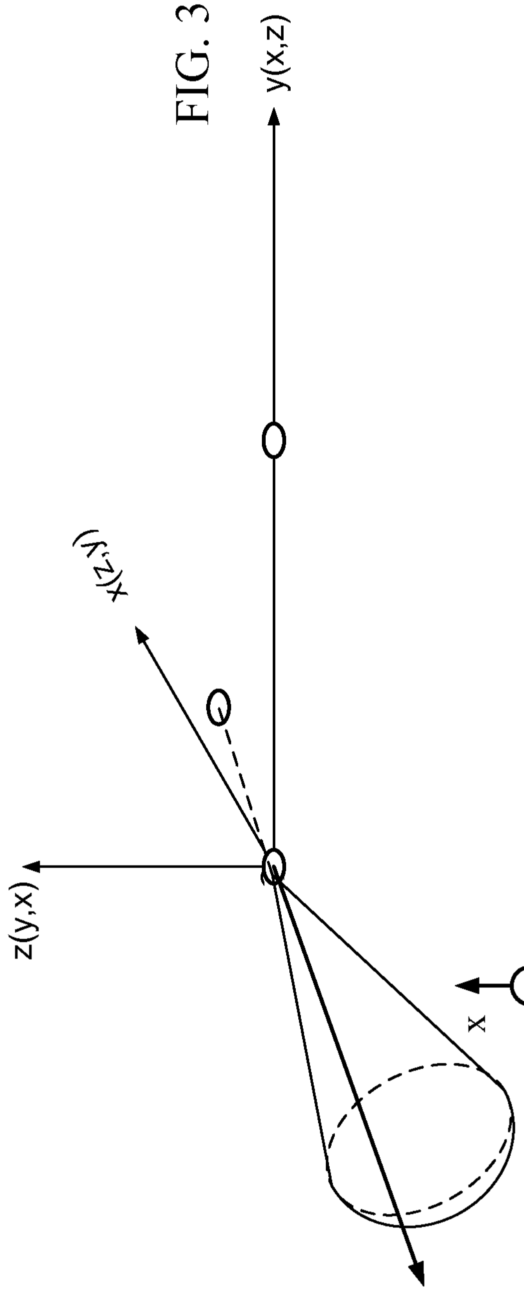


FIG. 34

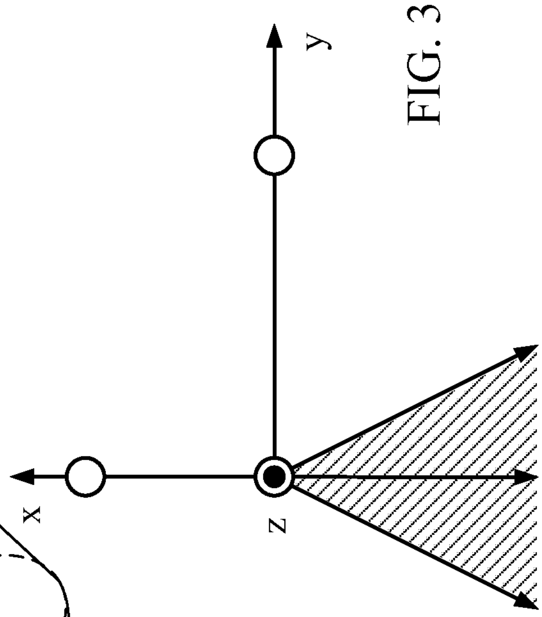


FIG. 35

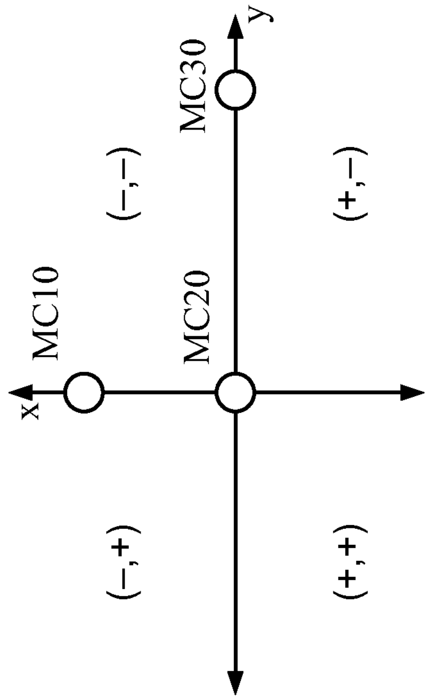


FIG. 36

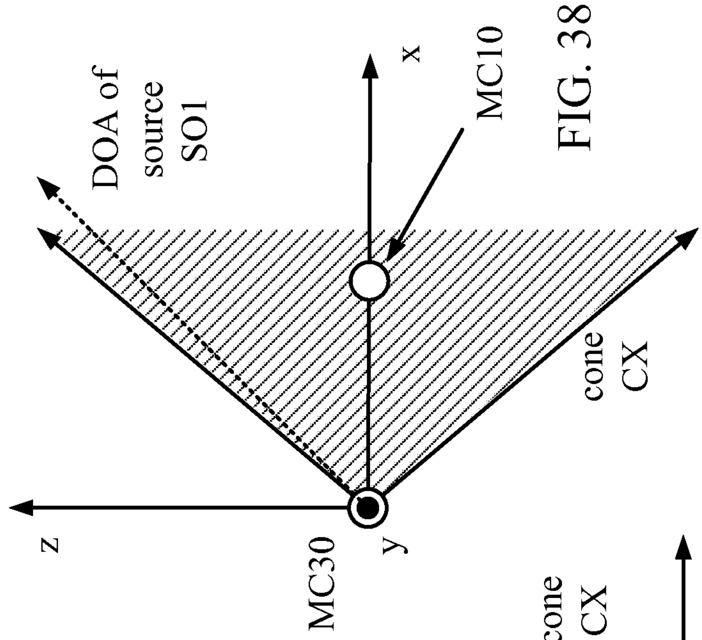


FIG. 38

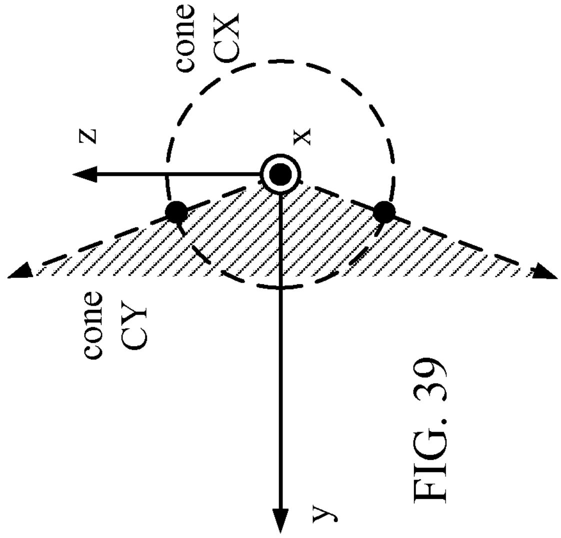


FIG. 39

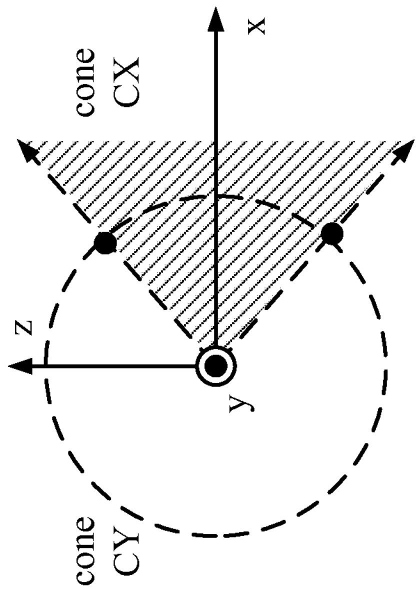


FIG. 40

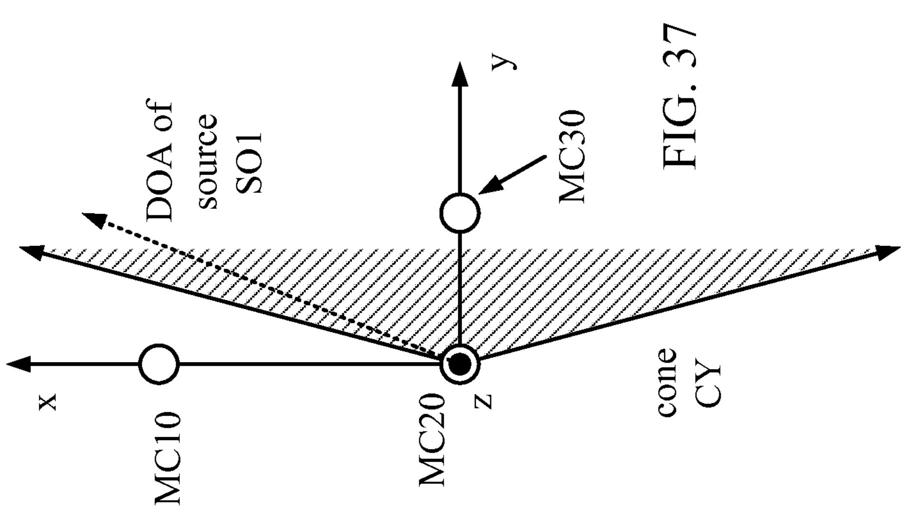


FIG. 37

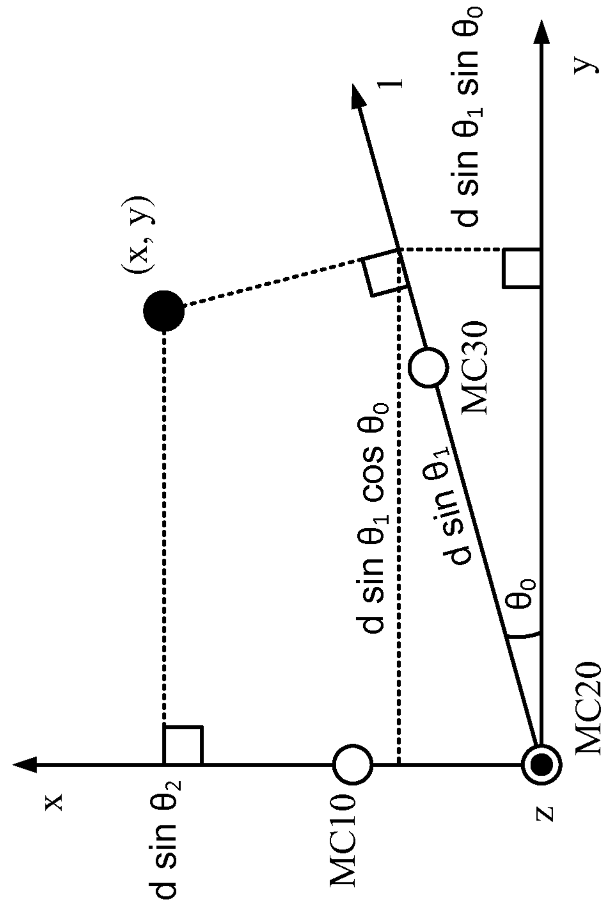

(x, y)
1
y
d sin θ1 sin θ0
d sin θ1 cos θ0
MC30
d sin θ1
θ0
MC20
x
d sin θ2
MC10
z

FIG. 42

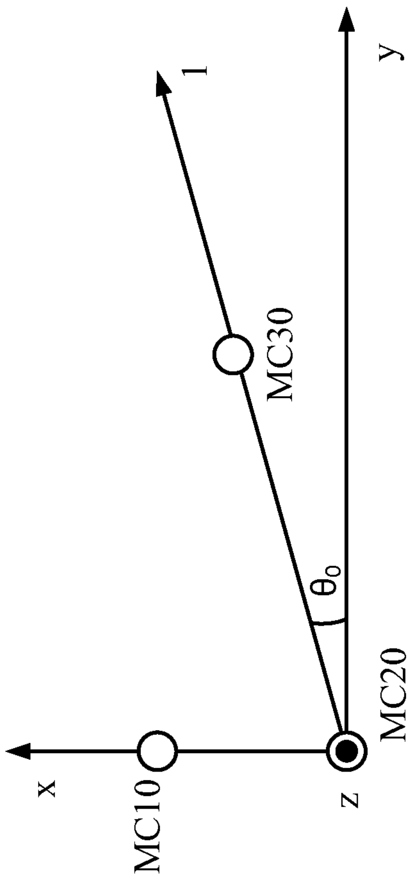

1
y
MC30
θ0
MC20
x
MC10
z

FIG. 41

steering vector

Target source number

pair number

mic number

$$\left(\begin{bmatrix} d^{1}_{1,1} & d^{2}_{1,1} & d^{3}_{1,1} \\ d^{1}_{1,2} & d^{2}_{1,2} & d^{3}_{1,2} \\ d^{1}_{2,1} & d^{2}_{2,1} & d^{3}_{2,1} \\ d^{1}_{2,2} & d^{2}_{2,2} & d^{3}_{2,2} \end{bmatrix}^{H} \begin{bmatrix} d^{1}_{1,1} & d^{2}_{1,1} & d^{3}_{1,1} \\ d^{1}_{1,2} & d^{2}_{1,2} & d^{3}_{1,2} \\ d^{1}_{2,1} & d^{2}_{2,1} & d^{3}_{2,1} \\ d^{1}_{2,2} & d^{2}_{2,2} & d^{3}_{2,2} \end{bmatrix} + \lambda \cdot I \right)^{-1} \begin{bmatrix} d^{1}_{1,1} & d^{2}_{1,1} & d^{3}_{1,1} \\ d^{1}_{1,2} & d^{2}_{1,2} & d^{3}_{1,2} \\ d^{1}_{2,1} & d^{2}_{2,1} & d^{3}_{2,1} \\ d^{1}_{2,2} & d^{2}_{2,2} & d^{3}_{2,2} \end{bmatrix}^{H} \begin{bmatrix} x_{1,1} \\ x_{1,2} \\ x_{2,1} \\ x_{2,2} \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix}$$

FIG. 43

$$\frac{\left(\begin{bmatrix} d^{1}_{1,1} & d^{2}_{1,1} & d^{3}_{1,1} \\ d^{1}_{1,2} & d^{2}_{1,2} & d^{3}_{1,2} \\ d^{1}_{2,1} & d^{2}_{2,1} & d^{3}_{2,1} \\ d^{1}_{2,2} & d^{2}_{2,2} & d^{3}_{2,2} \end{bmatrix}^{H} \begin{bmatrix} d^{1}_{1,1} & d^{2}_{1,1} & d^{3}_{1,1} \\ d^{1}_{1,2} & d^{2}_{1,2} & d^{3}_{1,2} \\ d^{1}_{2,1} & d^{2}_{2,1} & d^{3}_{2,1} \\ d^{1}_{2,2} & d^{2}_{2,2} & d^{3}_{2,2} \end{bmatrix} + \lambda \cdot I \right)^{-1} \begin{bmatrix} d^{1}_{1,1} & d^{2}_{1,1} & d^{3}_{1,1} \\ d^{1}_{1,2} & d^{2}_{1,2} & d^{3}_{1,2} \\ d^{1}_{2,1} & d^{2}_{2,1} & d^{3}_{2,1} \\ d^{1}_{2,2} & d^{2}_{2,2} & d^{3}_{2,2} \end{bmatrix}^{H} \begin{bmatrix} x_{1,1} \\ x_{1,2} \\ x_{2,1} \\ x_{2,2} \end{bmatrix}}{\underset{i}{\operatorname{argmax}} \left(\begin{bmatrix} d^{1}_{1,1} & d^{2}_{1,1} & d^{3}_{1,1} \\ d^{1}_{1,2} & d^{2}_{1,2} & d^{3}_{1,2} \\ d^{1}_{2,1} & d^{2}_{2,1} & d^{3}_{2,1} \\ d^{1}_{2,2} & d^{2}_{2,2} & d^{3}_{2,2} \end{bmatrix}^{H} \begin{bmatrix} d^{1}_{1,1} & d^{2}_{1,1} & d^{3}_{1,1} \\ d^{1}_{1,2} & d^{2}_{1,2} & d^{3}_{1,2} \\ d^{1}_{2,1} & d^{2}_{2,1} & d^{3}_{2,1} \\ d^{1}_{2,2} & d^{2}_{2,2} & d^{3}_{2,2} \end{bmatrix} + \lambda \cdot I \right)^{-1} \begin{bmatrix} d^{1}_{1,1} & d^{2}_{1,1} & d^{3}_{1,1} \\ d^{1}_{1,2} & d^{2}_{1,2} & d^{3}_{1,2} \\ d^{1}_{2,1} & d^{2}_{2,1} & d^{3}_{2,1} \\ d^{1}_{2,2} & d^{2}_{2,2} & d^{3}_{2,2} \end{bmatrix}^{H} \begin{bmatrix} d^{i}_{1,1} \\ d^{i}_{1,2} \\ d^{i}_{2,1} \\ d^{i}_{2,2} \end{bmatrix}} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix}$$

FIG. 44

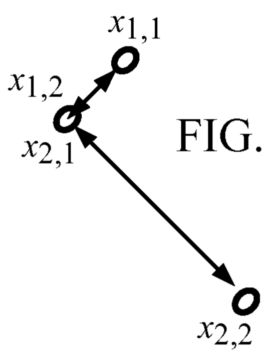
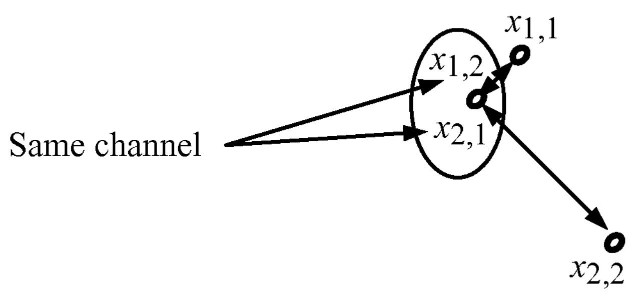


FIG. 45



$$\frac{\left(\begin{bmatrix} d^1_{1,1} & d^2_{1,1} & d^3_{1,1} \\ d^1_{1,2} & d^2_{1,2} & d^3_{1,2} \\ d^1_{2,2} & d^2_{2,2} & d^3_{2,2} \end{bmatrix}^H (\Gamma+\lambda\cdot I)^{-1} \begin{bmatrix} d^1_{1,1} & d^2_{1,1} & d^3_{1,1} \\ d^1_{1,2} & d^2_{1,2} & d^3_{1,2} \\ d^1_{2,2} & d^2_{2,2} & d^3_{2,2} \end{bmatrix}\right)^{-1} \begin{bmatrix} d^1_{1,1} & d^2_{1,1} & d^3_{1,1} \\ d^1_{1,2} & d^2_{1,2} & d^3_{1,2} \\ d^1_{2,2} & d^2_{2,2} & d^3_{2,2} \end{bmatrix}^H (\Gamma+\lambda\cdot I)^{-1} \begin{bmatrix} x_{1,1} \\ x_{1,2} \\ x_{2,2} \end{bmatrix}}{\underset{i}{\operatorname{argmax}} \left(\begin{bmatrix} d^1_{1,1} & d^2_{1,1} & d^3_{1,1} \\ d^1_{1,2} & d^2_{1,2} & d^3_{1,2} \\ d^1_{2,2} & d^2_{2,2} & d^3_{2,2} \end{bmatrix}^H (\Gamma+\lambda\cdot I)^{-1} \begin{bmatrix} d^1_{1,1} & d^2_{1,1} & d^3_{1,1} \\ d^1_{1,2} & d^2_{1,2} & d^3_{1,2} \\ d^1_{2,2} & d^2_{2,2} & d^3_{2,2} \end{bmatrix}\right)^{-1} \begin{bmatrix} d^1_{1,1} & d^2_{1,1} & d^3_{1,1} \\ d^1_{1,2} & d^2_{1,2} & d^3_{1,2} \\ d^1_{2,2} & d^2_{2,2} & d^3_{2,2} \end{bmatrix}^H (\Gamma+\lambda\cdot I)^{-1} \begin{bmatrix} d^i_{1,1} \\ d^i_{1,2} \\ d^i_{2,2} \end{bmatrix}}$$

FIG. 46

Target source number
1,2, ..., N, where N (number of sources) $\leq M$ (number of pairs)

steering vector
per pair

pair number

mic number

$$\left(\begin{bmatrix} d^1_{1,1} & \cdots & d^N_{1,1} \\ d^1_{1,2} & \cdots & d^N_{1,2} \\ d^1_{2,1} & \cdots & d^N_{2,1} \\ d^1_{2,2} & \cdots & d^N_{2,2} \\ \vdots & \ddots & \vdots \\ d^1_{M,1} & \cdots & d^N_{M,1} \\ d^1_{M,2} & \cdots & d^N_{M,2} \end{bmatrix}^H \begin{bmatrix} d^1_{1,1} & \cdots & d^N_{1,1} \\ d^1_{1,2} & \cdots & d^N_{1,2} \\ d^1_{2,1} & \cdots & d^N_{2,1} \\ d^1_{2,2} & \cdots & d^N_{2,2} \\ \vdots & \ddots & \vdots \\ d^1_{M,1} & \cdots & d^N_{M,1} \\ d^1_{M,2} & \cdots & d^N_{M,2} \end{bmatrix} + \lambda \cdot I \right)^{-1} \begin{bmatrix} d^1_{1,1} & \cdots & d^N_{1,1} \\ d^1_{1,2} & \cdots & d^N_{1,2} \\ d^1_{2,1} & \cdots & d^N_{2,1} \\ d^1_{2,2} & \cdots & d^N_{2,2} \\ \vdots & \ddots & \vdots \\ d^1_{M,1} & \cdots & d^N_{M,1} \\ d^1_{M,2} & \cdots & d^N_{M,2} \end{bmatrix}^H \begin{bmatrix} x_{1,1} \\ x_{1,2} \\ x_{2,1} \\ x_{2,2} \\ \vdots \\ x_{M,1} \\ x_{M,2} \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix}$$

FIG. 47

$$\left(d_{1,1}^{1}, d_{1,2}^{1}\right) = \left(\exp\left(-\frac{j2\pi f\, 0.073 \sin(\theta_1)}{c}\right), 1\right)$$

$$\left(d_{2,1}^{1}, d_{2,2}^{1}\right) = \left(\exp\left(-\frac{j2\pi f\, 0.036 \sin(\theta_2)}{c}\right), 1\right)$$

FIG. 48

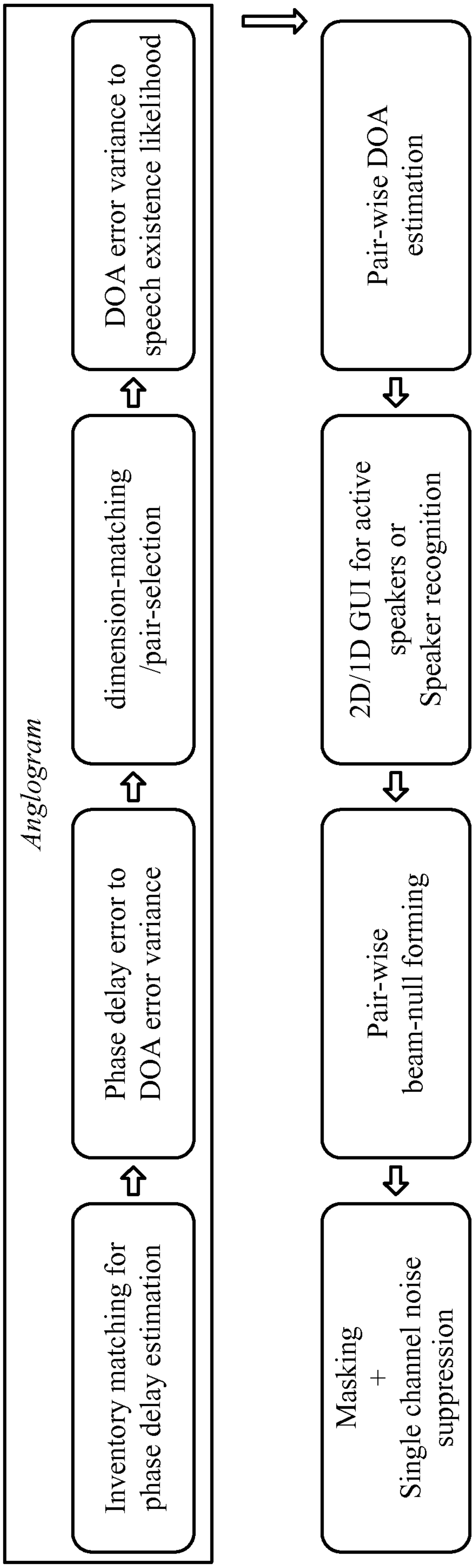

Anglogram
Inventory matching for phase delay estimation
Phase delay error to DOA error variance
dimension-matching /pair-selection
DOA error variance to speech existence likelihood
Pair-wise DOA estimation
2D/1D GUI for active speakers or Speaker recognition
Pair-wise beam-null forming
Masking + Single channel noise suppression

FIG. 49

INTERNATIONAL SEARCH REPORT

International application No
PCT/US2013/039624

A. CLASSIFICATION OF SUBJECT MATTER
INV. G10L25/48
ADD. G10L17/00 H04R1/40 G01S3/80 G10L25/06

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)
G04G G06F H04N G01S G10L H04M H04R H04L

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)

EPO-Internal, INSPEC, WPI Data

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	CANTON-FERRER C ET AL: "Multimodal real-time focus of attention estimation in SmartRooms", COMPUTER VISION AND PATTERN RECOGNITION WORKSHOPS, 2008. CVPR WORKSHOPS 2008. IEEE COMPUTER SOCIETY CONFERENCE ON, IEEE, PISCATAWAY, NJ, USA, 23 June 2008 (2008-06-23), pages 1-8, XP031285736, ISBN: 978-1-4244-2339-2 paragraphs [04.2], [0004], [0002], [0001] paragraph [0005] ----- -/--	1,17,28, 44,55, 67,78,94

[X] Further documents are listed in the continuation of Box C.

[X] 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 but 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
21 June 2013	04/07/2013

Name and mailing address of the ISA/	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Fax: (+31-70) 340-3016	Krembel, Luc

1

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	GB 2 342 802 A (PICTURETEL CORP [US]) 19 April 2000 (2000-04-19) figure 1 page 1, line 30 - page 2, line 5 page 6, line 23 - line 26 page 7, line 1 - line 9 page 8, line 10 - line 16 page 10, line 26 - page 11, line 24 -----	1,17,28, 44,55, 67,78,94
A	US 5 724 485 A (RAINTON DAVID [JP]) 3 March 1998 (1998-03-03) column 9, line 31 - column 10, line 57; figures 4,5 -----	1,17,28, 44,55, 67,78,94
A	US 2007/192103 A1 (SATO NOBUO [JP] ET AL) 16 August 2007 (2007-08-16) abstract paragraphs [0002], [0008] paragraph [0032] paragraphs [0035], [0036]; figure 2 paragraph [0041] figure 3 -----	1,17,28, 44,55, 67,78,94
A	HORI T ET AL: "Low-Latency Real-Time Meeting Recognition and Understanding Using Distant Microphones and Omni-Directional Camera", IEEE TRANSACTIONS ON AUDIO, SPEECH AND LANGUAGE PROCESSING, IEEE SERVICE CENTER, NEW YORK, NY, USA, vol. 20, no. 2, 1 February 2012 (2012-02-01), pages 499-513, XP011464135, ISSN: 1558-7916, DOI: 10.1109/TASL.2011.2164527 figure 3 paragraph [III.C] paragraph [0V.D] -----	1,17,28, 44,55, 67,78,94
A	HAYLEY HUNG ET AL: "Estimating Dominance in Multi-Party Meetings Using Speaker Diarization", IEEE TRANSACTIONS ON AUDIO, SPEECH AND LANGUAGE PROCESSING, IEEE SERVICE CENTER, NEW YORK, NY, USA, vol. 19, no. 4, 1 May 2011 (2011-05-01), pages 847-860, XP011352013, ISSN: 1558-7916, DOI: 10.1109/TASL.2010.2066267 paragraph [0V.A] ----- -/--	1,17,28, 44,55, 67,78,94

1

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	WO 2006/006935 A1 (AGENCY SCIENCE TECH & RES [SG]; CHEN JIANFENG [SG]; XU YONG [SG]; LIU) 19 January 2006 (2006-01-19) abstract figure 1	1,17,28, 44,55, 67,78,94

1

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No
PCT/US2013/039624

Patent document cited in search report		Publication date	Patent family member(s)		Publication date
GB 2342802	A	19-04-2000	GB 2342802	A	19-04-2000
			JP 2000125274	A	28-04-2000
US 5724485	A	03-03-1998	JP 2749780	B2	13-05-1998
			JP H08101693	A	16-04-1996
			US 5724485	A	03-03-1998
US 2007192103	A1	16-08-2007	JP 5055781	B2	24-10-2012
			JP 2007218933	A	30-08-2007
			US 2007192103	A1	16-08-2007
			US 2012004915	A1	05-01-2012
WO 2006006935	A1	19-01-2006	NONE		

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US2013/039624

Box No. II Observations where certain claims were found unsearchable (Continuation of item 2 of first sheet)

This international search report has not been established in respect of certain claims under Article 17(2)(a) for the following reasons:

1. ☐ Claims Nos.:
because they relate to subject matter not required to be searched by this Authority, namely:

2. ☒ Claims Nos.:
because they relate to parts of the international application that do not comply with the prescribed requirements to such an extent that no meaningful international search can be carried out, specifically:
see FURTHER INFORMATION sheet PCT/ISA/210

3. ☐ Claims Nos.:
because they are dependent claims and are not drafted in accordance with the second and third sentences of Rule 6.4(a).

Box No. III Observations where unity of invention is lacking (Continuation of item 3 of first sheet)

This International Searching Authority found multiple inventions in this international application, as follows:

1. ☐ As all required additional search fees were timely paid by the applicant, this international search report covers all searchable claims.

2. ☐ As all searchable claims could be searched without effort justifying an additional fees, this Authority did not invite payment of additional fees.

3. ☐ As only some of the required additional search fees were timely paid by the applicant, this international search report covers only those claims for which fees were paid, specifically claims Nos.:

4. ☐ No required additional search fees were timely paid by the applicant. Consequently, this international search report is restricted to the invention first mentioned in the claims; it is covered by claims Nos.:

Remark on Protest

☐ The additional search fees were accompanied by the applicant's protest and, where applicable, the payment of a protest fee.

☐ The additional search fees were accompanied by the applicant's protest but the applicable protest fee was not paid within the time limit specified in the invitation.

☐ No protest accompanied the payment of additional search fees.

FURTHER INFORMATION CONTINUED FROM PCT/ISA/ 210

Continuation of Box II.2

Claims Nos.: 2-16, 29-43, 56-66, 79-93(completely); 1, 28, 55, 78(partially)

Claims 1,28,55,78 do not meet the requirements of Article 6 PCT to such an extent that no meaningful search is possible.
The following discussion quotes features of claim 1. The same reasoning applies mutatis mutandis to claims 28,55,78.
The feature "determining the social interaction between participants based on the similarity between the first spatially filtered output and each of the second spatially filtered outputs" used in these claims does not have a clearly defined technical meaning in the sense of Rule 6.3(b) PCT. The term "social interaction" as such does not have a well defined meaning, even while considering the indications of paragraphs [0060],[0075], figures 4-7. It is in particular insufficiently disclosed in the sense of Article 5 PCT, how to determine any "social interaction" feature illustrated in figures 4-7, using a similarity value. It is therefore not possible to meaningfully search the feature of "determining the social interaction (...)". The term "beamformer" is ambiguous since it may be applied to different spatial filtering techniques, for example to sound waves or to radio waves (antennas, radar). The present application does only support beamformer applied to microphone arrays. The definition "a fixed beamformer (...) configured to receive a plurality of second spatially filtered outputs from a plurality of steerable beamformers " is not supported by the description in the sense of Article 6 PCT. As far as can be understood from figure 3 (or alternatively figures 18,20) and claim 2, the "fixed beamformer" claimed comprises/is associated to a fixed array of microphones ("black dots" in figure 3) and comprises possibly also the units ref. 320, 325. On the other hand, the "steerable beamformers" comprise a plurality of microphone arrays processed by units ref.305. Figure 3 does not support the definition "a fixed beamformer (...) configured to receive a plurality of second spatially filtered outputs from a plurality of steerable beamformers". Only unit ref.340 receives the outputs of said plurality of steerable beamformers. Since the definition "(...) configured to receive (...)" is not supported by the application, it appears not to make sense to carry out a search for this feature. It is therefore not possible to carry out a meaningful search for the above indicated independent claims. Since no meaningful search is possible for the above mentioned independent claims, no search has been performed for their dependent claims, because these dependent claims do not further clarify the subject-matter to which they refer.

The applicant's attention is drawn to the fact that claims relating to inventions in respect of which no international search report has been established need not be the subject of an international preliminary examination (Rule 66.1(e) PCT). The applicant is advised that the EPO policy when acting as an International Preliminary Examining Authority is normally not to carry out a preliminary examination on matter which has not been searched. This is the case irrespective of whether or not the claims are amended following receipt of the search report or during any Chapter II procedure. If the application proceeds into the regional phase before the EPO, the applicant is reminded that a search may be carried

FURTHER INFORMATION CONTINUED FROM PCT/ISA/ 210

out during examination before the EPO (see EPO Guidelines C-IV, 7.2), should the problems which led to the Article 17(2) declaration be overcome.