(19) Octrooicentrum Nederland

(11) **2021308**

(12) B1 **OCTROOI**

(21) Aanvraagnummer: **2021308**

(22) Aanvraag ingediend: **16 juli 2018**

(51) Int. Cl.:
G10L 21/0216 (2018.01) **H04R 3/00** (2018.01) **G10L 15/26** (2019.01)

(30) Voorrang:

(41) Aanvraag ingeschreven:
24 januari 2020

(43) Aanvraag gepubliceerd:
-

(47) Octrooi verleend:
24 januari 2020

(45) Octrooischrift uitgegeven:
30 januari 2020

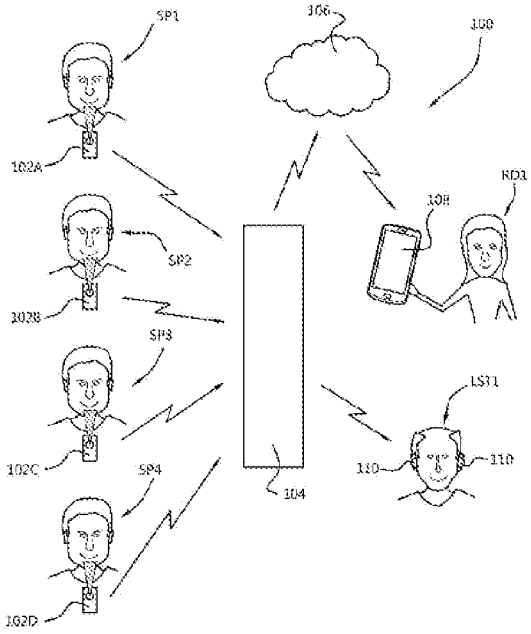
(73) Octrooihouder(s):
Hazelebach & van der Ven Holding B.V.
te Rotterdam

(72) Uitvinder(s):
Marcellus Wilhelmus Jacobus van der Ven
te Rotterdam
Jari Tamas Hazelebach te Rotterdam

(74) Gemachtigde:
ir. J.H. de Hoog te Veenendaal

(54) **METHODS FOR A VOICE PROCESSING SYSTEM**

(57) Methods for a voice processing system comprising P microphone units (102A...102D) and a central unit (104) are disclosed. Each microphone unit is linked to a person and derives from N microphone signals a source localisation signal. The source localisation signal is used to control an adaptive beam form process to obtain a beam formed audio signal. The microphone unit is further configured to derive metadata from for N microphone signals, such direction the sound is coming from. Packages with the metadata and beam formed audio signal are transmitted to the central unit. The central unit processes the metadata to determine which parts of the P beam formed audio signal comprises speech from a person that is linked to another microphone unit. By removing said parts from the audio signals before transcription, the quality of the transcription is improved. The transcriptions are displayed on a remote device.



NL B1 2021308

METHODS FOR A VOICE PROCESSING SYSTEM

TECHNICAL FIELD

The subject disclosure relates to the field of voice processing. More particular the subject disclosure relates to a method in a microphone unit of a voice processing system and a method in a central unit of a voice processing system. Furthermore, the subject disclosure relates to a microphone unit and a central unit of a voice processing comprising P microphone units.

BACKGROUND ART

EP3057340A1 discloses a hearing system comprising a partner microphone unit. The partner microphone comprises a multitude of microphones for picking up a target signal component, i.e. the person's voice, and a noise component, i.e. background sound. From the microphone signals a beam-formed signal is generated, wherein signal components from other directions than a direction of the target signal are attenuated, whereas signal components from the directions of the target signal source are left un-attenuated. The partner microphone comprises further antenna and transceiver circuitry for establishing a wireless audio link to another device, e.g. a hearing aid, headset.

US2002/0150263A1 discloses a signal processing system which receives signals from a number of different sensors which are representative of signals generated from a plurality of sources. The sensed signals are processed to determine the relative position of each of the sources relative to the sensors. This information is then used to separate the signals from each of the sources. The system can be used, for example, to separate the speech signal generated from a number of users in a meeting.

US2015/0154183A1 discloses an auto-translation conferencing system for multi user audio. A translation services server receives over a separate communication channel the audio signals from different devices for translation into textual data. The textual data may be translated into text of different languages based on the language preferences of the end user devices.

In the known systems multiple microphones at fixed positions are used to separate the speech signal generated from a number of users in e.g. a meeting. If the speech signals could not be separate with enough quality, a

speech signal assumed to comprise only voice of a first speaker could also comprise voice of other speakers. When the first speaker is listening to another speaker, the voice of another speaker could become dominant in the speech signal. When the speech signal is supplied to a translation service, the translated speech signal comprises both text spoken by the first speaker and the other speaker. Furthermore, when the users are not at a fixed position and could move around in the room, it would be hard to obtain a good speech signal of the person moving around.

SUMMARY OF INVENTION

It is an object of the present technology to provide a voice processing system for processing simultaneously voice of multiple persons in a conversation which has at least one of the advantages over the known voice processing systems: improved sound quality of the individual speaker, not limited to speakers at fixed positions, improved signals to be supplied to translation services, reduced crosstalk of speakers in audio channel of a speaker, flexible in use, scalable with respect to number of users, simultaneous translation of more than one conversation in a single room, reduced power consumption.

According to the subject technology, this object is achieved by a method in a microphone unit having the features of claim 1 and a method in a central unit having the features of claim 11. Advantageous embodiments and further ways of carrying out the present technology may be attained by the measures mentioned in the dependent claims.

According to a first aspect of the subject technology, there is provided a method in a microphone unit of a voice processing system comprising P microphone units and a central unit. The microphone unit retrieves from N input units Mic_i, i=1, 2, ...N, N≥2, N microphone signals having a first sampling frequency SF1, each microphone signal comprising a target signal component and a noise signal component. The unit determines from the N microphone signals a source localisation signal having a second sampling frequency SF2, wherein SF1≥SF2. The unit derives from a group of Y consecutive samples of the source localisation signal a beam form control signal. Under control of the beam form control signal the unit generates a group of Y consecutive samples of a beam formed audio signal having a sampling frequency SF2 from the N microphone

signals. The unit derives a set of metadata for the group of Y consecutive samples of the beam formed audio signal from corresponding samples of the N microphone signals from which the group of Y consecutive samples of the beam formed audio signal has been obtained. The unit further generates data packages and streams wirelessly the data packages to the central unit of the voice processing system. Each data package comprises Q groups of Y consecutive samples of the beam formed audio signal and Q sets of metadata derived for Q groups of Y consecutive samples from corresponding samples of the N microphone signals.

There is further provided a method in the central unit. The central unit receives wirelessly P streams of data packages from P microphone units. Each data package comprises Q groups of Y consecutive samples of a beam formed audio signal and Q sets of metadata corresponding to Q groups of Y consecutive samples of the beam formed audio signal. The central unit time synchronizes the data packages of the P streams to obtain P synchronized streams of data packages. The central unit detects in each of the P synchronized streams based on the beam formed audio signals and time corresponding metadata which parts of the P beam formed audio signals comprises a target signal component of an active speaker linked to the microphone unit which generated said stream and forwards the detected parts of the beam formed audio signals of the P streams for further processing.

The present technology is based on the insight that for a good transcription of speech of a conversation into text, it is important that each person taking part of the conversation is clearly identified in and isolated from the captured audio signals by the microphones of the voice processing system. The term ‘microphone unit’ in the subject technology is to be understood in relation to a user wearing the ‘microphone unit’ and which speech has to be processed for reproduction and/or transcription. The microphone unit is preferably attached to the person. Even though the microphone unit is placed relatively close to the sound source of interest (the mouth of the wearer), the target-signal-to-noise ratio of the signal picked up by the microphone may still be less than desired, for example due to back ground noise or other persons who speak very loudly. Beam forming is used to improve the target-signal-to-noise ratio of each microphone unit. However, it is still possible that the voice of a person who speaks very loudly has

enough amplitude in the beam formed audio signal transmitted by the microphone unit such that it is recognized as speech by a transcription engine. In that case, the transcription of the audio signal from the microphone unit results in text coming from the user wearing the microphone unit and at least one other loudly speaking person. By generating set of metadata from the N microphones of a microphone unit which provides information about the assumed audio source linked to the microphone unit which could not be derived from the beam formed audio signal from said microphone unit, the central unit will be able to combine content of sets of the meta data from all microphone units to improve detection of which parts of the received beam formed audio signal from the P microphone units comprises speech of the person associated with the microphone unit and which parts probably do not comprises speech of said person. According to the present technology, the metadata comprises at least information that is derived from a combination of the N microphones signals of the microphone unit and which could not be derived from the beam formed audio signal at the output of the microphone unit. For example a field of a set of metadata could indicate from which direction the sound is coming from or could indicate whether more than a predefined percentage of Y consecutive samples is coming from a direction falling within a range defined by used the used beam-shaping algorithm. For example, when the metadata indicates that a part of the beam formed audio signal from a particular microphone unit probably does not comprise target sound, the central unit could verify whether the metadata of any other time corresponding parts of the received beam formed audio signals from the other microphone units comprises target sound. If this is the case, it is very likely that said part of the audio signal from the particular microphone unit does not comprise speech from the person associated with said microphone unit, as a general rule in conversations is that only one person is speaking at a time. This identified part could subsequently be removed or attenuated before for example transcription of the audio signal.

In an embodiment, a value of a first metadata field is derived from a group of Y consecutive samples of the source localisation signal. Normally, a microphone unit attached to the cloth of a person will receive sound from said person from a particular direction resulting in a corresponding value of the source localisation signal. When the sound is coming from another direction, the source localisation signal will have a value significantly differing from the value in case the

linked person is speaking. This is an indication that the sound is from another source and that it should probably be removed from the audio stream before transcription to text. The central unit could use this information to start a verification process to verify whether another microphone unit had received earlier the same speech of said person. If that is the case, the corresponding part of the beamed formed audio signal could be removed before transcription of the speech into text.

In an embodiment of the subject technology, the sets of metadata and Q groups of Y consecutive samples of the beam formed audio signal that have been derived from a corresponding part in time of the N microphone signals are included in a i^{th} data package and $i+T^{th}$ data package respectively, wherein T is an integer greater than 0. In this embodiment the metadata is transmitted some time in advance of the corresponding beam formed audio signal. It has been found that for a good transcription of speech it is advantageous to have some milliseconds of background noise before the speech in the audio signal. If the first word of speech starts with a plosive phoneme, for example the phoneme of the letters p, t, k, b and d in Dutch and said phonemes is not preceded by background noise, said letter is regularly missed in the text of the transcript of the audio signal, which is not the case when having some background noise in advance of a plosive phoneme. Furthermore, the central unit could benefit from these features as it could start analysis of the metadata of other audio signals in advance of receiving the part of the audio signal comprising speech and will therefore be able to minimize the throughput time from receiving the beam formed audio signal and forwarding the audio signal for further processing.

In an embodiment of the subject disclosure, a sample of the source localisation signal has a value indicating the direction from which is estimated that the target signal component is coming from; the microphone unit determines the number of samples from the Y consecutive samples of the source localisation signal that have a value in a range defined by the beam form control signal. If the number is larger than a predefined threshold value, the microphone unit inserts in a set of the metadata a second field with a value indicating the corresponding Y consecutive samples of the beam formed audio signal comprises target sound. In another embodiment, the number is included in a field of a set of metadata.

In an embodiment of the subject disclosure, streaming of data packages is started when the metadata of a package indicates that the time corresponding Y consecutive samples of the beam formed audio signal comprises target sound. This feature enables to reduce the power consumption of the microphone unit by switching on the transceiver at moments that the audio signal is expected to comprise speech. In a further embodiment, streaming of data packages is stopped after at least T data packages comprising metadata indicating that the corresponding Q groups of Y consecutive samples of the beam formed audio signal does not comprise target sound. This feature improves the quality of the transcription and enables to reduce the power consumption of the microphone unit and thereby extend the operating time on one battery charge.

In an embodiment of the subject disclosure, the central unit generates for each of the P microphone units a streaming control signal based on the beam formed audio signals and time corresponding metadata and transmits the streaming control signals to the P microphone units. The microphone unit receives from the central unit a streaming control signal and stops streaming data packages in response to the streaming control signal. This feature enables to further reduce power consumption of a microphone unit.

In an embodiment of the subject disclosure, a set of metadata comprises for each of the Y consecutive samples of the beam formed audio signal a field having a value derived from the corresponding samples of the source localisation signal. Having the direction from which the audio is coming from for each group of Y consecutive samples, enables the central unit to improve the decision whether a group of Y consecutive samples of a microphone unit comprises speech of the person associated with said microphone unit.

In an embodiment, the method in the microphone unit determines a speaker voice profile from the N microphone signals and verifies if the speaker voice profile corresponds to a microphone reference speaker voice profile. When the speaker voice profile corresponds to the microphone reference speaker voice profile, the microphone unit start streaming the data packages. In a further embodiment, the microphone reference speaker voice profile is received from the central unit. These features enables to reduces the power consumption for transmitting packages by starting only when the voice in the beam formed audio signal is likely to come from the speaker wearing the microphone unit.

According to a second aspect of the subject technology there is provided a microphone unit and a central unit having a processor and a memory to store instructions that, when executed by the processor, cause the microphone unit and central unit to perform corresponding methods descript above.

Other features and advantages will become apparent from the following detailed description, taken in conjunction with the accompanying drawings which illustrate, by way of example, various features of embodiments.

BRIEF DESCRIPTION OF THE DRAWINGS

These and other aspects, properties and advantages will be explained hereinafter based on the following description with reference to the drawings, wherein like reference numerals denote like or comparable parts, and in which:

Fig. 1 shows schematically a voice processing system according to the present technology;

Fig. 2 shows schematically an embodiment of a microphone unit;

Fig. 3 shows schematically an embodiment of a central unit.

DESCRIPTION OF EMBODIMENTS

The advantages, and other features of the technology disclosed herein, will become more readily apparent to those having ordinary skill in the art from the following detailed description of certain preferred embodiments taken in conjunction with the drawings which set forth representative embodiments of the present technology.

Figure 1 shows schematically an embodiment of a voice processing system 100 according the present subject technology. The voice processing system comprises a multitude of microphone unit's 102A-102D, a central unit 104, a transcription server in the cloud 106, a mobile device 108 and a hearing aid 110. Each microphone unit 102A-102D is attached to a speaking person SP1-SP4 taking part of a conversation. A microphone unit is configured for picking up sound and generating an audio signal to be wirelessly transmitted to the central unit 104. The audio signal comprises a target signal component and a noise

signal component, wherein the target signal component is coming from the speaking person that wears the microphone unit and the noise component is all the other sound in the audio signal, e.g. ambient noise and sound of loud speaking person. Placing a wireless microphone close to a sound source of interest makes communication in challenging environments easier. The microphone unit transforms the sound captured by its microphones in to a digital audio signal that can be wirelessly transmitted to the central unit 104. Any digital transmission protocol can be used to communicate wirelessly with the central unit. Examples of transmission protocols are, but not limited to: WIFI, DECT, and Bluetooth.

The central unit 104 receives the digital audio signals from the P microphone unit 102A-102D connected to the central unit 104. The central unit 104 simultaneously processes the digital audio signals to remove speech parts from a digital audio signal from a microphone unit which are not coming from the person using said microphone unit and optionally to remove some noise from each of the P digital audio signals to obtain P quality improved digital audio signals. The P quality improved digital audio signals are transmitted over the internet to a transcription service 106. The transcription service 106 is configured to transcript the P quality improved digital audio signals into text and to send said text to one or more remote display devices 108, such as but not limited to: mobile phones, tablets, laptops, desktops, smart TVs. An app running on the remote display devices 108 is configured to display the text as a conversation on it screen to enable a reader RD1 to follow the conversation by reading the text.

An app running on the remote display devices 108 enables a user to link the device to a specific central unit. An on-line communication service handles user requests to link the central unit and routes the text obtained by transcription of the P quality improved audio signals corresponding to the P microphone units coupled to the central unit via the internet to the remote display devices linked to said central unit. The transcription service 106 and communication service may run on the same server or on different servers in the cloud.

Optionally, the central unit 104 is configured to combine the P quality improved digital audio signals to a combined audio signal for transmission to one or more head phones or hearing aids 110 worn by a person LST1 with for example a hearing impairment. With this embodiment, the speech of each person taking

part of a conversation is optimized to be reproduced by the hearing aid. For example, the loudness of the different speaking persons may be equalized and/or the frequency spectrum of each person may be adapted such that a person with a hearing impairment is capable to follow the conversation by listening.

Fig. 2 shows schematically an embodiment of a microphone unit 200. The microphone unit 200 comprises N microphones Mic_1 ... Mic_N, a processor 210 and a transceiver 226. A microphone is a transducer that converts sound into an electrical signal. In the present embodiment each microphone generates a PDM signal with a sample frequency SF1 of 2048 kHz. PDM or Pulse dense modulation is a form of modulation used to represent an analog signal with a binary signal. The PDM signals are supplied to inputs of the processor 210. The N microphones are omnidirectional microphones located at some distance from each other to enable detection from which direction the audio is coming from.

The processor 210 receives the PDM signals. A conversion function 212 converts the 1-bit PDM signals from the N microphones to corresponding N 16-bit Pulse-code modulation (PCM) signals with a sampling frequency SF2 of 16 kHz. The conversion unit 212 comprises a decimation filter to enable down sampling of the audio signals without losing information in the audio signals to detect from which direction target sound is coming from.

A source localisation function 214 derives a source localisation signal from the N 16-bit PCM signals obtained from the conversion unit 212. A value of the source localisation signal calculated from 16 subsequent samples the N 16 bit PCM signals and corresponds to the direction the main signal component in the 16 subsequent samples is coming from. The source localisation signal has a sampling frequency SF3 of 1 kHz. In another embodiment, the source localisation signal includes a signal indicating the direction of the target sound coming from and the distance between the target source and the microphone unit.

Block 216 an algorithm to calculate a beam form control signal from the source localisation signal. In an embodiment, the algorithm performs a low-pass filter function to obtain a relative slowly changing beam form control signal. In another embodiment, the algorithm determine from a specified number of subsequent samples of the source localisation the direction which occurs most often and uses this direction to generate the corresponding value of the beam form control signal.

The beam form control signal controls an adaptive beam filter process 218. In an embodiment, the adaptive beam filter process 218 is configured to select two of the N microphone signals under control of the beam form control signal to obtain a 16 bit beam formed audio signal having a sampling frequency of 16kHz from said two microphone signals. Said two microphone signals correspond to the two microphones having the largest distance from each other in the direction of the target sound indicated by the beam form control signal. As the distance between the two selected microphones is known, the assumed direction of the target sound and the speed of sound in air, it is possible to combine the two audio signals such that the target sound is amplified and noise is attenuated. When the distance between target source and microphone unit and direction of target sound are used by the adaptive beam filter the signal quality of the beam formed audio signal could be improved. It should be noted that beam forming is a commonly known technology. As the beam formed audio signal will be submitted to a transcription engine, the adaptive beam filter process performs optionally a normalisation filter. Normalisation of a speech signal improves the quality of the transcription process as the loudness of each speech part will be substantially the same.

The 16 kHz samples of the beam formed audio signal are supplied to a First-In First-Out (FIFO) buffer 220. The FIFO buffer 220 has a buffer size of Y× T samples, wherein Y and T are integers with a value larger than 0.

Block 216 further represents the generation of a set of metadata for a group of Y subsequent samples of the beam formed audio signal. The set of metadata is derived from corresponding samples of the N microphone signal from which the group of Y consecutive samples of the beam formed audio signal has been obtained. In an embodiment, a first metadata field is derived from a group of Y consecutive samples of the source localisation signal. In an embodiment, Y = 16. This means that one first metadata field is generated each millisecond. In an embodiment, a sample of the source localisation signal has a value indicating the direction from which is estimated that the target signal component is coming from and block 216 determines the number of samples from the Y consecutive samples of the source localisation signal that have a value in a range defined by the beam form control signal. If the number is larger than a predefined threshold value, block 216 generates a set of metadata with a second field with a value indicating

that the corresponding Y consecutive samples of the beam formed audio signal comprises target sound. How this metadata is used in the central unit 104 will be described below.

Block 222 represents a function which generates data packages comprising Q groups of Y consecutive samples of the beam formed audio signal and Q sets of metadata, wherein a set of metadata is obtained for a group of Y consecutive samples of the beam formed audio signal. Due to the buffer 220 having a size of Y×T samples, an i^{th} data package comprises a set of meta data that have been derived from a part of the microphone signals Mic with a length of Y/SF2 second which is T×Y/SF2 seconds in time before the part of the microphone signals that has been used to obtain the group of Y consecutive samples of the beam formed audio signal. In other words: the Q sets of metadata and Q groups of Y consecutive samples of the beam formed audio signal that have been derived from substantially the same part in time of the N microphone signals are included in a i^{th} data package and $i+T^{th}$ data package respectively. An advantage that the sets of metadata derived from a part in time of the N microphone signals Mic_1 … Mic_N arrives at the packaging function some time before the correspond group of Y consecutive samples of the beam formed audio derived from the same part in time of the N microphone signals is that when starting streaming of the data packages due to the detection of voice, the stream of data packages starts with a defined minimal number of groups of Y consecutive samples of the beam formed audio signal which comprises no voice. It has been found when transcription speech, the transcription improves when the speech is preceded by a time period without speech. The buffer 220 enables this.

Block 224 controls the streaming of the data packages to the central unit. If the metadata indicates for a predetermined time that the beam formed audio signal does not comprise speech, the streaming of data packages is stopped to reduce power consumption by bringing the transceiver 226 in low power mode. In an embodiment, as soon as the metadata indicates that the coming groups of Y consecutive samples of the beam formed audio signal comprises speech, the transceiver will become in transmission mode and will start transmission of the data packages. Transceiver 226, e.g. a WIFI transceiver, is configured to stream wirelessly the data packages MU_x to the central unit 104. Index x indicates the index number of the microphone unit. In an embodiment, an

internet protocol is used to communicate with the central unit. To reduce overhead in the communication, UDP (User Datagram Protocol) can be used. It might be clear to the skilled person that each communication protocol that has enough bandwidth to transmit the data packages might be used.

Optionally, the transceiver 226 is configured to receive a microphone control signal MCS_x from the central unit. In an embodiment, the microphone control signal comprises a stop command. When the function in block 224 detects the stop command, the streaming of data packages is stopped and the transceiver 226 is switched in low power mode. In an embodiment, the microphone control signal further carries a reference speaker voice profile. The reference speaker voice profile for a microphone unit is obtained by analysing the beam formed audio signal from said microphone unit which is transmitted to a transcription engine. The mel-frequency cepstrum (MFC) is a possible format for a speaker profile. However any other suitable format could be used to characterize voice of a person.

In an embodiment, the microphone unit comprises functionality to determine from the beam formed audio signal or microphone signals Mic_1 ... Mic_N an actual speaker profile. The microphone unit is further configured to verify whether the actual speaker profile corresponds to the reference speaker voice profile retrieved from the microphone control signal MCS_X received from the central unit. If there is a more than a predefined degree of similarity between the actual speaker profile and the reference speaker voice profile, the transmission of data packages could be started by the microphone unit. In this way, when the central unit detects by processing the P streams of data packages from the P microphone units that the stream of microphone unit MU_x comprises speech from a person other than the person wearing the microphone unit, the central unit stops streaming of data packages by a microphone unit MU_x by sending the stop signal in the Microphone Control Signal MCS_x to said microphone unit MU_x. As soon as the Microphone unit MU_x detects that the reference speaker voice profile has a predefined degree of similarity, said microphone unit MU_x will start streaming of the data packages again. In this way, the power consumption of the microphone unit could be reduced further.

Block 215 performs one or more functions to derive special characteristics from the N PDM signals coming from the microphones. The

special characteristics have a property that could not accurately be derived from the beam formed audio signal that will be transmitted to the central unit for further processing and transcription. Examples of such special characteristics are not limited to: maximum Signal to Noise Ratio (SNR) of the microphone signals Mic_1 – Mic_N, signal power (dB) of the microphone signals Mic_1 – Mic_N, etc.. The derived special characteristics are transmitted in fields of a set of metadata to the central unit. The derived special characteristics enables the central unit to determine more accurately which parts of the streamed beam formed audio signal from a particular microphone unit do not comprise speech from the speaker associated with said particular microphone. By removing these parts before transcription, the quality of the transcription could be improved.

Fig. 3 shows schematically an embodiment of a central unit 300. The central unit 300 comprises a first transceiver 320, a signal processor 310 and a second transceiver 360. The first transceiver 320, e.g. a WIFI receiver, is configured to communicate simultaneously with the P microphone units 102A …102D of the voice processing system 100 in Fig. 1.. The signals from the P microphones units MU_1 … MU_P comprises for each microphone unit MU_i a stream of data packages. The stream of data packages may be a continuous stream of data packages or non-continuous stream of data packages. The transceiver forwards the P streams to a synchronization process 330. As the internet protocol is used for communication, the arrival time at the central unit 300 of a data package from a first microphone unit MU_1 corresponding to a particular moment in time does not necessarily have the same arrival time at the central unit of a data package from a second microphone MU_2. The synchronization process 330 uses time stamps in the data packages to align the streams of data packages in time and forwards P time synchronized streams from the microphone units to an automatic speech detection process 340. Technology for time synchronizing data is commonly known to the skilled person and therefor this is not descripted in more detail.

The automatic speech detection process 340 is configured to process the time synchronized sets of metadata of the P microphone units. As described above the metadata of a microphone unit MU_i comprises at least one data field with characteristics or features derived from the N microphone signals Mic_1 … Mic_N which could not be derived from the beam formed audio signal of

said microphone unit MU_1. An example of such a characteristic is the determined angle from which the target sound is coming from. Another example is a field indicating the minimal Signal to Noise Ration of a part of the N microphone signals time corresponding to a group of Y consecutive sample of the beam formed audio signal. Another example is a field with a value indicating that the corresponding Y consecutive samples of the beam formed audio signal comprises target sound which value is derived from the source localisation signal. In still another embodiment, the last or first sample of the source localisation signal corresponding a group of Y consecutive beam formed audio samples is stored in a data field of a set of metadata.

By combining the content of the set of metadata it is possible to determine that some parts of the streamed beam formed audio signal do not comprise speech of the person to which the microphone is assigned. For example, normally as a rule in a conversation one person is speaking at a time. However, when one person changes his voice from a normal volume to a loud volume, next to the microphone unit that is worn by said person, another microphone unit could pick up his voice and could incorrectly conclude that this is voice from the person wearing the other microphone and starts streaming the speech of the loud speaking person to the central unit. As the sound is coming from another direction than the target sound is normally coming from, the central unit could detect this in the metadata of the another microphone and as normally as a rule not two persons are speaking at the same time, the automatic speech detection process will detect this constellation and will remove the corresponding data packages from the stream of the other microphone to obtain for each microphone unit is reduced stream of data packages. The decision algorithm could be improved by taking into account known or calculated distance between the microphone and the speed of sound in air. Similarly, the signal power corresponding to a group of Y consecutive samples of the beam formed audio signal and Signal to noise ratio might be used to improve detection whether or not the audio in a data package in the stream of a particular microphone unit MU_i comprise speech from the person associated with said microphone unit.

In an embodiment, when the automatic speech detection process 340 detects that a data stream from a microphone unit carries speech from another person wearing another microphone unit, the process 340 generates a

stop command for said microphone unit and supplies the stop command to the transceiver 320 for transmission to the corresponding microphone unit. In response to the stop command, the microphone unit switches its transceiver in low power mode to reduce power consumption. The stop command is submitted as part of the microphone control signal. The automatic speech detection process 340 supplies the P reduced streams of data packages to a further processing process 350. In an embodiment, the further processing process 350 is configured to determine for each of the P microphone unit from the beam formed audio signal in the data packages of the P reduced streams a reference speaker voice profile. The reference speaker voice profile for microphone unit MU_i is supplied to the transceiver 320 for submission to the corresponding microphone unit MU_i as part of the microphone control signal MSC_i. The microphone unit MU_i might use the reference speaker voice profile to compare a speaker profile derived from the N microphone PDM signals. If there is enough similarity, the streaming control process 224, activates the transceiver into transmission mode and starts transmission of the data packages.

The further processing process 350 is further configured to stream the beam formed audio signals of the reduced stream of the P microphone units to an in the cloud multi speaker transcription engine via transceiver 360 with an internet protocol. In an embodiment the transceiver 360 uses an UDP protocol. The multi speaker transcription engine is configured to receive the beam formed audio signals of the reduced streams, to transcript separately each of the beam formed audio signals into text and to transmit the text such that the text can be displayed on a display of a remote device as a conversation wherein on the display is indicated for each part of the text an indication of the audio source. An indication of the audio source might be a microphone index number, name assigned to a microphone unit, unique name for each speaker, etc. In an embodiment a reader RD1 could get the transcription by accessing a webpage via a unique URL assigned to the conversation. New text of transcribed speech will be pushed to all devices that are linked to said transcription by accessing the URL. The transcribed speech can be displayed as webpage or by an app running on the device.

The further processing process 350 is optionally configured to combine the beam formed audio signals of the reduced stream of the P

microphone units to a combined audio signal OUT2 that can be transmitted to an audio system, headphone or hearing aid 110 such that people LST1 can follow the conversation by listening. In this embodiment, the parts of the beam formed audio signal that have been removed by the automatic speech detection process will not be included in the combined audio signal. In this way echo cancelation is applied.

A microphone unit for use in a voice processing system described above comprises comprises N microphones generating N microphone signals having a first sampling frequency SF1, a wireless communication unit, a processor and a memory to store instructions that, when executed by the processor, cause the microphone unit to perform any of the methods for a microphone unit described above.

A central unit for use in a voice processing system described above comprises a wireless communication unit configured to communicate with each of the P microphone units, a processor and a memory to store instructions that, when executed by the processor, cause the central unit to perform any of the method for a microphone unit described above.

It should be noted that it might be possible to include the transcription engine function in the central unit 104. This would make the system a stand-alone system. However this requires a lot of local processing power making the system unnecessary expensive.

In the subject disclosure, sample frequencies of 2048 kHz, 16 kHz and 1 kHz have been used as exemplary embodiments. The use of other sample frequencies will not change the concept of the subject disclosure.

The present subject disclosure describes a computer implemented method in a microphone unit and a central unit. The central unit might be in the form of a dedicated hardware unit with a processor comprising instructions when executed cause the processor to perform the method. However, the central unit might also be implemented in a mobile device such as but not limited to: smartphone, tablet and laptop.

The subject disclosure transcribing a conversation between two or more persons in the same room is very useful for persons with a hearing loss, that are hearing impaired, deaf and/or hard-of-hearing. They could follow the conversation by reading. The subject disclosure can also be applied for transcribing meetings and could replace a note taker, secretary or translator. The

latter is possible when automatically a translation is made of the transcribed text from the spoken language to another language.

While the invention has been described in terms of several embodiments, it is contemplated that alternatives, modifications, permutations and equivalents thereof will become apparent to those skilled in the art upon reading the specification and upon study of the drawings. The invention is not limited to the illustrated embodiments. Changes can be made without departing from the scope which is defined by the appended claims.

-O-O-O-O-O-

Conclusies:

1. Werkwijze in een microfooneenheid van een spraakverwerkingssysteem omvattende P-microfooneenheden en een centrale eenheid, waarbij de werkwijze omvat:

- het ophalen uit N invoereenheden Mic_i, i=1, 2, ...N, N≥2, van N microfoonsignalen met een eerste bemonsteringsfrequentie SF1, waarbij elk microfoonsignaal een doelsignaalcomponent en een ruissignaalcomponent omvat;
- het bepalen uit de N-microfoon van een bronlocaliseringssignaal met een tweede bemonsteringsfrequentie SF2, waarbij SF1≥SF2;
- het afleiden van een bundelvormbesturingssignaal uit een groep van Y opeenvolgende monsters van het bronlocaliseringssignaal;
- het onder besturing van het bundelvormbesturingssignaal genereren uit de N-microfoonsignalen van een groep van Y opeenvolgende monsters van een bundelgevormd audiosignaal met een bemonsteringsfrequentie SF2;
- het afleiden van een set van metadata voor de groep van Y opeenvolgende monsters van het bundel gevormde audiosignaal uit overeenkomstige monsters van de N-microfoonsignalen waaruit de groep van Y opeenvolgende monsters van het bundelgevormde audiosignaal is verkregen;
- het genereren van datapakketten, waarbij een datapakket omvat Q-groepen van Y opeenvolgende monsters van het bundelgevormde audiosignaal en Q-sets van metadata, waarbij Q≥1;
- het draadloos streamen van de datapakketten naar de centrale eenheid.

2. De werkwijze volgens conclusie 1, waarbij een waarde van een eerste metadataveld wordt afgeleid van een groep Y opeenvolgende monsters van het bronlocaliseringssignaal.

3. De werkwijze volgens willekeurig welke van de conclusies 1 - 2, waarbij de sets van metadata en Q-groepen van Y opeenvolgende monsters van het bundelgevormde audiosignaal die zijn afgeleid van een overeenkomstig deel in de tijd van de N-microfoonsignalen respectievelijk zijn opgenomen in een i^{th}. datapakket en i + T^{th} datapakket , waarbij T een geheel getal groter dan 0 is.

4. De werkwijze volgens willekeurig welke van de conclusies 1 - 3, waarbij een monster van het bronlocaliseringssignaal een waarde heeft die de richting aangeeft van waaruit wordt geschat dat de doelsignaalcomponent afkomstig is; waarbij de werkwijze verder omvat:

- het bepalen van het aantal monsters van de Y opeenvolgende monsters van het bronlocaliseringssignaal die een waarde hebben in een bereik gedefinieerd door het bundelvormbesturingssignaal;
- en als het aantal groter is dan een vooraf gedefinieerde drempelwaarde, het invoegen in een tweede veld van de metadata met een waarde die aangeeft dat de corresponderende Y opeenvolgende monsters van het bundel gevormde audiosignaal doelgeluid bevatten.

5. De werkwijze volgens conclusie 4, waarbij het streamen van datapakketten wordt gestart wanneer een set van de metadata van een pakket aangeeft dat de corresponderende Y opeenvolgende monsters van het bundel gevormde audiosignaal doelgeluid bevatten.

6. De werkwijze volgens conclusie 5 in combinatie met conclusie 3, waarbij het streamen van datapakketten wordt gestopt nadat ten minste T-datapakketten metadata omvatten waarbij het tweede veld van die ten minste T-datapakketten aangeeft dat de corresponderende Y opeenvolgende samples van het bundelgevormde audiosignaal geen doelsignaal omvatten.

7. De werkwijze volgens willekeurig welke van de conclusies 1 - 6, waarbij de werkwijze omvat:

het bepalen van een sprekerspraakprofiel uit de N-microfoonsignalen;
het controleren of het sprekerspraakprofiel overeenkomt met een spraakprofiel van een referentiesprekerspraakprofiel van een microfoon; en
het beginnen met het streamen van de datapakketten wanneer het sprekerspraakprofiel overeenkomt met het referentiesprekerspraakprofiel van de microfoon.

8. De werkwijze volgens willekeurig welke van de conclusies 1 - 7, waarbij de werkwijze verder omvat:

- het ontvangen van het referentiesprekerspraakprofiel van de microfoon via de centrale eenheid.

9. De werkwijze volgens willekeurig welke van de conclusies 1 - 8, waarbij de werkwijze verder omvat:
- het ontvangen van een microfoonbesturingssignaal via de centrale eenheid;
- het stoppen met streamen van datapakketten in reactie op het microfoonbesturingssignaal.

10. De werkwijze volgens willekeurig welke van de conclusies 1 - 9, waarbij de metadata voor elk van de Y opeenvolgende monsters van het bundelgevormde audiosignaal een derde veld met een waarde afgeleid van de overeenkomstige monsters van het bronlocaliseringssignaal omvat.

11. Werkwijze in een centrale eenheid van een spraakverwerkingssysteem omvattende P-microfooneenheden die de werkwijze volgens willekeurig welke van de conclusies 1 - 10 uitvoeren, waarbij de werkwijze omvat:
- het draadloos ontvangen van P-stromen met datapakketten van de P-microfooneenheden, elk datapakket omvat Q-groepen van Y opeenvolgende monsters van een bundelgevormd audiosignaal en Q sets van metadata corresponderend met Q-groepen van Y opeenvolgende monsters van het bundelgevormde audiosignaal;
- het in tijd synchroniseren van de datapakketten van de P-streams om P-gesynchroniseerde stromen van datapakketten te verkrijgen;
- het detecteren in elk van de P-gesynchroniseerde stromen welke delen van de bundel gevormde audiosignalen een doelsignaalcomponent omvatten van een actieve spreker die gelinkt is met de microfooneenheid die genoemde stroom genereerde op basis van de bundelgevormde audiosignalen en met de in tijd overeenkomende sets van metadata;
- het doorsturen van de gedetecteerde delen van de bundelgevormde audiosignalen van de P-stromen voor verdere verwerking.

12. De werkwijze volgens conclusie 11, waarbij de metadata in het i^{th}-datapakket overeenkomt met het bundelgevormde audiosignaal in het i + T^{th}-datapakket, waarbij T een geheel getal groter dan 0 is.

13. De werkwijze volgens willekeurig welke van de conclusies 11 - 12, verder omvattende:
- het genereren van een microfoonbesturingssignaal voor elk van de P-microfooneenheden op basis van de bundelgevormde audiosignalen en in tijd corresponderende sets van metadata;
- het verzenden van de microfoonbesturingssignalen naar de P-microfooneenheden.

14. De werkwijze volgens willekeurig welke van de conclusies 11 - 13, verder omvattende:
- het verkrijgen van een sprekerspraakprofiel voor elk van de P-microfooneenheden;
- het verzenden van het sprekerspraakprofiel van microfooneenheid MU_x naar microfooneenheid MU_x als het referentiesprekerspraakprofiel.

15. De werkwijze volgens willekeurig welke van de conclusies 1 - 14, waarbij de set van metadata ten minste één veld omvat dat een karakteristiek weergeeft, genomen uit een groep, omvattende: doelbronlocatie, SNR, DB, spraak gedetecteerd.

16. Microfooneenheid voor gebruik in een spraakverwerkingssysteem omvattende P-microfooneenheden en een centrale eenheid, de microfooneenheid omvat N-microfoons die microfoonsignalen genereren met een eerste bemonsteringsfrequentie SF1, een draadloze communicatie-eenheid, een processor en een geheugen om instructies op te slaan die, wanneer uitgevoerd door de processor, bewerkstelligen dat de microfooneenheid de werkwijze volgens willekeurig welke van de conclusies 1 – 10 uitvoert.

17. Centrale eenheid voor gebruik in een spraakverwerkingssysteem omvattende P-microfooneenheden en een centrale eenheid, de centrale eenheid

omvat een draadloze communicatie-eenheid geconfigureerd om te communiceren met elk van de P-microfooneenheden, een processor en een geheugen om instructies op te slaan die, wanneer uitgevoerd door de processor, bewerkstelligen dat de centrale eenheid de werkwijze volgens willekeurig welke van de conclusies 11 - 15 uitvoert.

–○–○–○–○–○–

Fig. 1

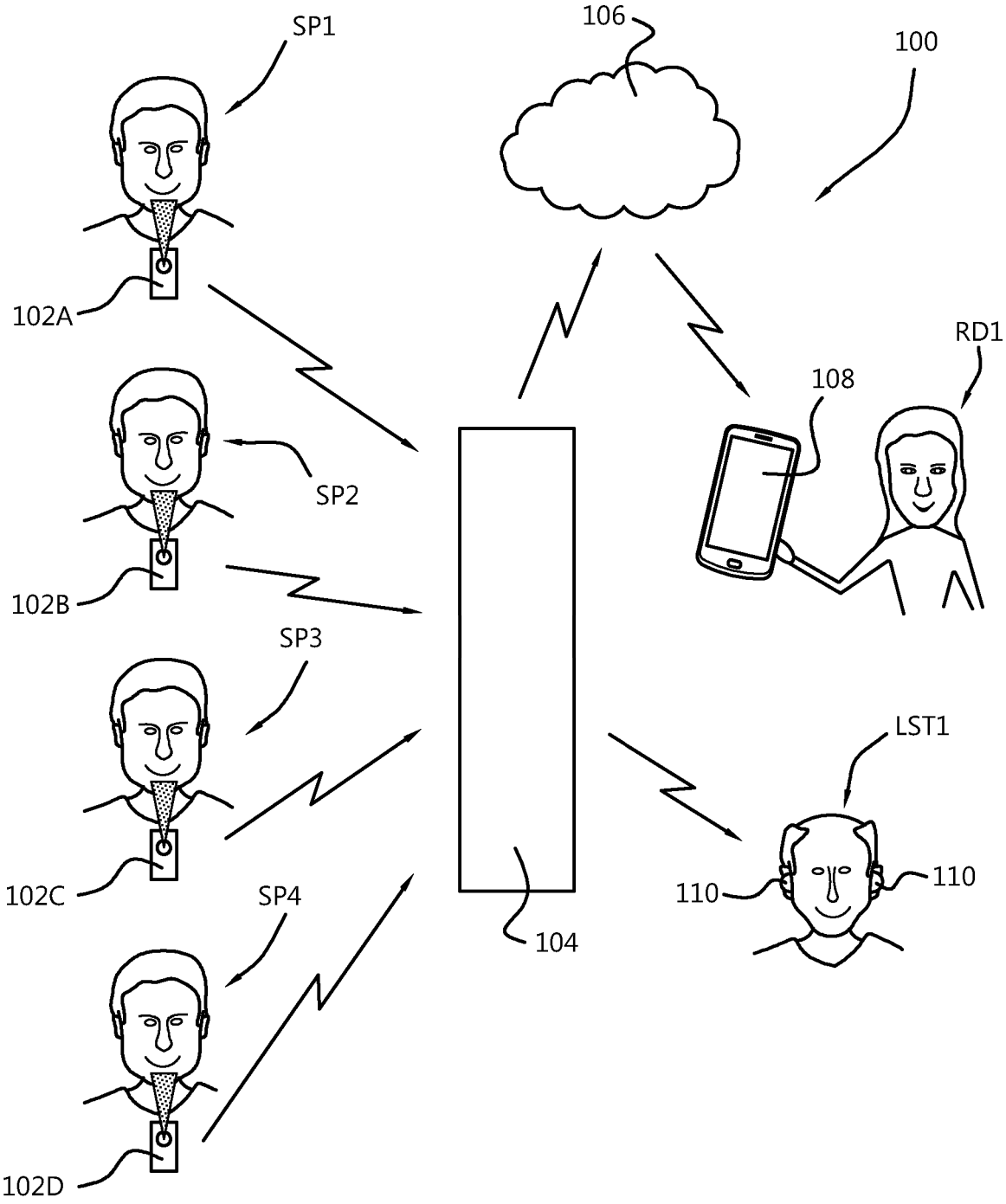

SP1
106
100
102A
RD1
108
SP2
102B
SP3
LST1
102C
110
110
SP4
104
102D

Fig. 2

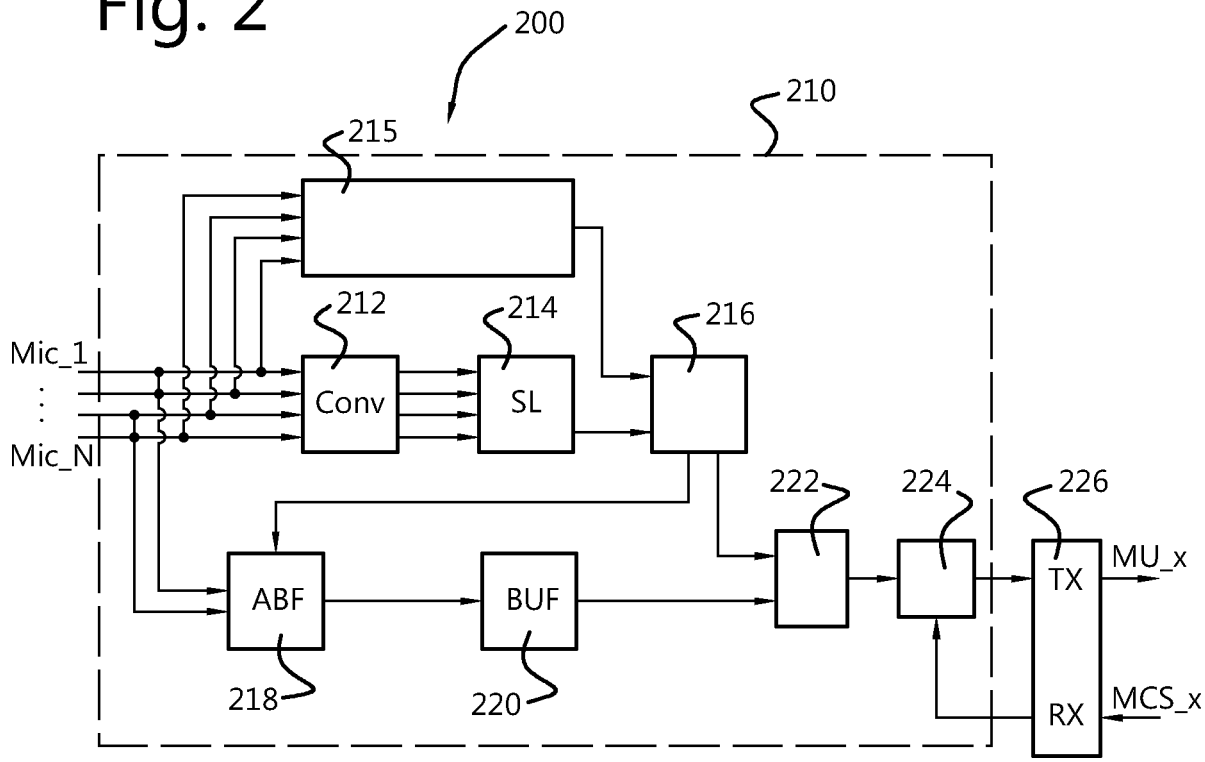
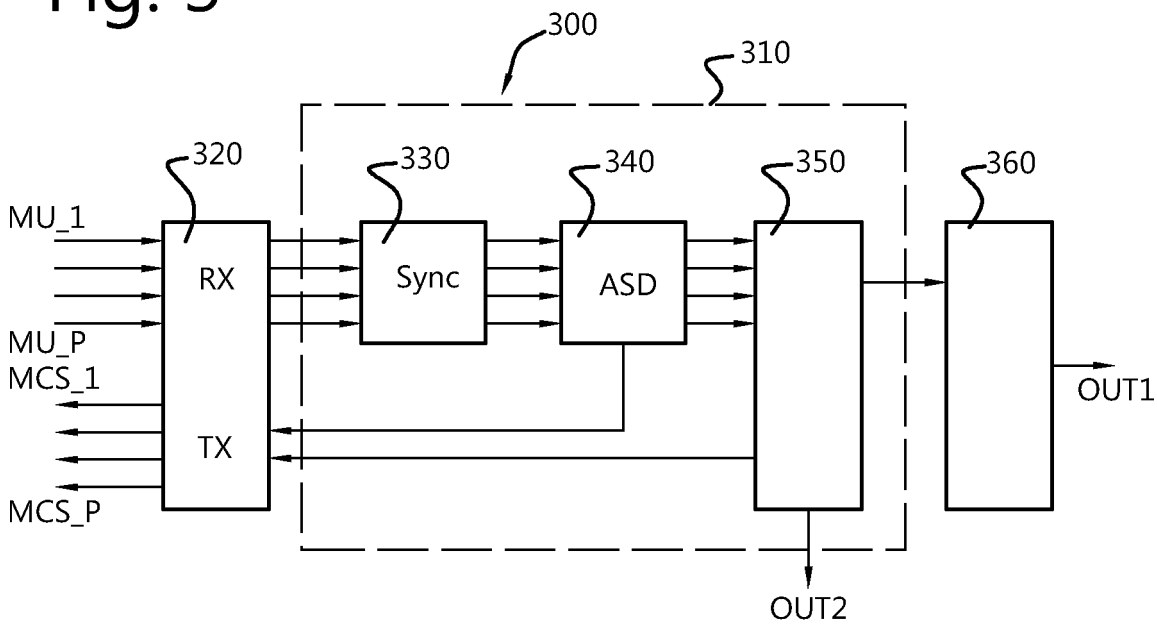

200
210
215
212
214
216
Mic_1
Mic_N
Conv
SL
222
224
226
ABF
BUF
TX
MU_x
RX
MCS_x
218
220

Fig. 3


300
310
320
330
340
350
360
MU_1
MU_P
MCS_1
MCS_P
RX
TX
Sync
ASD
OUT1
OUT2

ABSTRACT

Methods for a voice processing system comprising P microphone units (102A...102D) and a central unit (104) are disclosed. Each microphone unit is linked to a person and derives from N microphone signals a source localisation signal. The source localisation signal is used to control an adaptive beam form process to obtain a beam formed audio signal. The microphone unit is further configured to derive metadata from for N microphone signals, such direction the sound is coming from. Packages with the metadata and beam formed audio signal are transmitted to the central unit. The central unit processes the metadata to determine which parts of the P beam formed audio signal comprises speech from a person that is linked to another microphone unit. By removing said parts from the audio signals before transcription, the quality of the transcription is improved. The transcriptions are displayed on a remote device.

Fig. 1

SAMENWERKINGSVERDRAG (PCT)

RAPPORT BETREFFENDE NIEUWHEIDSONDERZOEK VAN INTERNATIONAAL TYPE

IDENTIFICATIE VAN DE NATIONALE AANVRAGE	KENMERK VAN DE AANVRAGER OF VAN DE GEMACHTIGDE
	0125P-NL
Nederlands aanvraag nr. **2021308**	Indieningsdatum **16-07-2018**
	Ingeroepen voorrangsdatum
Aanvrager (Naam) **Hazelebach & van der Ven Holding B.V.**	
Datum van het verzoek voor een onderzoek van internationaal type **18-09-2018**	Door de Instantie voor Internationaal Onderzoek aan het verzoek voor een onderzoek van internationaal type toegekend nr. **SN71991**

I. CLASSIFICATIE VAN HET ONDERWERP (bij toepassing van verschillende classificaties, alle classificatiesymbolen opgeven)

Volgens de internationale classificatie (IPC)

H04R3/00;G10L21/0216;G10L15/26

II. ONDERZOCHTE GEBIEDEN VAN DE TECHNIEK

Onderzochte minimumdocumentatie

Classificatiesysteem	Classificatiesymbolen
IPC	**G10L;H04R**

Onderzochte andere documentatie dan de minimum documentatie, voor zover dergelijke documenten in de onderzochte gebieden zijn opgenomen

III. ☐	**GEEN ONDERZOEK MOGELIJK VOOR BEPAALDE CONCLUSIES**	(opmerkingen op aanvullingsblad)
IV. ☐	**GEBREK AAN EENHEID VAN UITVINDING**	(opmerkingen op aanvullingsblad)

ONDERZOEKSRAPPORT BETREFFENDE HET RESULTAAT VAN HET ONDERZOEK NAAR DE STAND VAN DE TECHNIEK VAN HET INTERNATIONALE TYPE

Nummer van het verzoek om een onderzoek naar de stand van de techniek
NL 2021308

A. CLASSIFICATIE VAN HET ONDERWERP
INV. H04R3/00
ADD. G10L21/0216 G10L15/26

Volgens de Internationale Classificatie van octrooien (IPC) of zowel volgens de nationale classificatie als volgens de IPC

B. ONDERZOCHTE GEBIEDEN VAN DE TECHNIEK

Onderzochte minimum documentatie (classificatie gevolgd door classificatiesymbolen)
G10L H04R

Onderzochte andere documentatie dan de minimum documentatie, voor dergelijke documenten, voor zover dergelijke documenten in de onderzochte gebieden zijn opgenomen

Tijdens het onderzoek geraadpleegde elektronische gegevensbestanden (naam van de gegevensbestanden en, waar uitvoerbaar, gebruikte trefwoorden)
EPO-Internal, WPI Data

C. VAN BELANG GEACHTE DOCUMENTEN

Categorie °	Geciteerde documenten, eventueel met aanduiding van speciaal van belang zijnde passages	Van belang voor conclusie nr.
X A	US 2013/024196 A1 (GANONG III WILLIAM F [US] ET AL) 24 januari 2013 (2013-01-24) * bladzijde 2, alinea 20 - bladzijde 10, alinea 102; figuren 1, 3A, 4 * -----	1-12, 14-17 13
A	US 2013/166299 A1 (SHIMOTANI KEI [JP] ET AL) 27 juni 2013 (2013-06-27) * bladzijde 1, alinea 22 - bladzijde 2, alinea 30; figuren 1, 3, 9-10, 12, 14 * -----	1-17
A	WO 2015/008162 A2 (VOCAVU SOLUTIONS LTD [IL]) 22 januari 2015 (2015-01-22) * bladzijde 5, alinea 13 - bladzijde 11, alinea 28; figuren 1, 3 * -----	1-17

-/--

[X] Verdere documenten worden vermeld in het vervolg van vak C.

[X] Leden van dezelfde octrooifamilie zijn vermeld in een bijlage

° Speciale categorieën van aangehaalde documenten

"A" niet tot de categorie X of Y behorende literatuur die de stand van de techniek beschrijft

"D" in de octrooiaanvrage vermeld

"E" eerdere octrooi(aanvrage), gepubliceerd op of na de indieningsdatum, waarin dezelfde uitvinding wordt beschreven

"L" om andere redenen vermelde literatuur

"O" niet-schriftelijke stand van de techniek

"P" tussen de voorrangsdatum en de indieningsdatum gepubliceerde literatuur

"T" na de indieningsdatum of de voorrangsdatum gepubliceerde literatuur die niet bezwarend is voor de octrooiaanvrage, maar wordt vermeld ter verheldering van de theorie of het principe dat ten grondslag ligt aan de uitvinding

"X" de conclusie wordt als niet nieuw of niet inventief beschouwd ten opzichte van deze literatuur

"Y" de conclusie wordt als niet inventief beschouwd ten opzichte van de combinatie van deze literatuur met andere geciteerde literatuur van dezelfde categorie, waarbij de combinatie voor de vakman voor de hand liggend wordt geacht

"&" lid van dezelfde octrooifamilie of overeenkomstige octrooipublicatie

Datum waarop het onderzoek naar de stand van de techniek van internationaal type werd voltooid	Verzenddatum van het rapport van het onderzoek naar de stand van de techniek van internationaal type
26 maart 2019	

Naam en adres van de instantie	De bevoegde ambtenaar
European Patent Office, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Fax: (+31-70) 340-3016	Duffner, Orla

1

ONDERZOEKSRAPPORT BETREFFENDE HET RESULTAAT VAN HET ONDERZOEK NAAR DE STAND VAN DE TECHNIEK VAN HET INTERNATIONALE TYPE

Nummer van het verzoek om een onderzoek naar de stand van de techniek
NL 2021308

C.(Vervolg) VAN BELANG GEACHTE DOCUMENTEN

Categorie *	Geciteerde documenten, eventueel met aanduiding van speciaal van belang zijnde passages	Van belang voor conclusie nr.
A	EP 3 057 340 A1 (OTICON AS [DK]) 17 augustus 2016 (2016-08-17) * kolom 2, alinea 10 - kolom 20, alinea 91; figuren 1A, 4-5 * -----	1-17
A	US 8 144 903 B2 (PHONAK AG) 27 maart 2012 (2012-03-27) * kolom 1, regel 7 - kolom 7, regel 65 * -----	1-17

ONDERZOEKSRAPPORT BETREFFENDE HET RESULTAAT VAN HET ONDERZOEK NAAR DE STAND VAN DE TECHNIEK VAN HET INTERNATIONALE TYPE

Informatie over leden van dezelfde octrooifamilie

Nummer van het verzoek om een onderzoek naar de stand van de techniek
NL 2021308

In het rapport genoemd octrooigeschrift	Datum van publicatie	Overeenkomend(e) geschrift(en)	Datum van publicatie
US 2013024196 A1	24-01-2013	GEEN	
US 2013166299 A1	27-06-2013	JP 5867066 B2	24-02-2016
		JP 2013134312 A	08-07-2013
		US 2013166299 A1	27-06-2013
WO 2015008162 A2	22-01-2015	US 2016179831 A1	23-06-2016
		WO 2015008162 A2	22-01-2015
EP 3057340 A1	17-08-2016	CN 105898662 A	24-08-2016
		EP 3057340 A1	17-08-2016
		US 2016241975 A1	18-08-2016
		US 2018048969 A1	15-02-2018
US 8144903 B2	27-03-2012	EP 2103175 A1	23-09-2009
		EP 2408222 A1	18-01-2012
		US 2010135512 A1	03-06-2010
		WO 2008074350 A1	26-06-2008

WRITTEN OPINION

File No.	Filing date *(day/month/year)*	Priority date *(day/month/year)*	Application No.
SN71991	16.07.2018		NL2021308

International Patent Classification (IPC)
INV. H04R3/00 ADD. G10L21/0216 G10L15/26

Applicant
Hazelebach & van der Ven Holding B.V.

This opinion contains indications relating to the following items:

- ☒ Box No. I Basis of the opinion
- ☐ Box No. II Priority
- ☐ Box No. III Non-establishment of opinion with regard to novelty, inventive step and industrial applicability
- ☐ Box No. IV Lack of unity of invention
- ☒ Box No. V Reasoned statement with regard to novelty, inventive step or industrial applicability; citations and explanations supporting such statement
- ☐ Box No. VI Certain documents cited
- ☐ Box No. VII Certain defects in the application
- ☐ Box No. VIII Certain observations on the application

	Examiner
	Duffner, Orla

Application number

WRITTEN OPINION

NL2021308

Box No. I Basis of this opinion

1. This opinion has been established on the basis of the latest set of claims filed before the start of the search.

2. With regard to any **nucleotide and/or amino acid sequence** disclosed in the application and necessary to the claimed invention, this opinion has been established on the basis of:

   a. type of material:

   ☐ a sequence listing

   ☐ table(s) related to the sequence listing

   b. format of material:

   ☐ on paper

   ☐ in electronic form

   c. time of filing/furnishing:

   ☐ contained in the application as filed.

   ☐ filed together with the application in electronic form.

   ☐ furnished subsequently for the purposes of search.

3. ☐ In addition, in the case that more than one version or copy of a sequence listing and/or table relating thereto has been filed or furnished, the required statements that the information in the subsequent or additional copies is identical to that in the application as filed or does not go beyond the application as filed, as appropriate, were furnished.

4. Additional comments:

Box No. V Reasoned statement with regard to novelty, inventive step or industrial applicability; citations and explanations supporting such statement

1. Statement

Novelty	Yes:	Claims	1-17
	No:	Claims	
Inventive step	Yes:	Claims	13
	No:	Claims	1-12, 14-17
Industrial applicability	Yes:	Claims	1-17
	No:	Claims	

2. Citations and explanations

see separate sheet

Re Item V

Reasoned statement with regard to novelty, inventive step or industrial applicability; citations and explanations supporting such statement

1 Reference is made to the following documents:

D1 US 2013/166299 A1 (SHIMOTANI KEI [JP] ET AL) 27 juni 2013 (2013-06-27)

D2 WO 2015/008162 A2 (VOCAVU SOLUTIONS LTD [IL]) 22 januari 2015 (2015-01-22)

D3 EP 3 057 340 A1 (OTICON AS [DK]) 17 augustus 2016 (2016-08-17)

D4 US 2013/024196 A1 (GANONG III WILLIAM F [US] ET AL) 24 januari 2013 (2013-01-24)

D5 US 8 144 903 B2 (HAENGGI STEFAN [CH]; MARQUIS FRANCOIS [CH]; PLATZ RAINER [CH]; PHONAK AG [CH]) 27 maart 2012 (2012-03-27)

2 **Lack of inventive step, claim 1**

The present application does not meet the criteria of patentability, because the subject-matter of claim 1 does not involve an inventive step.

2.1 D1 is considered to be the closest prior art with respect to independent claim 1 and describes a method in a microphone unit (one of multiple devices 110A, 110B, 110C, 110D) of a speech processing system, the method comprising:

- retrieving from N input units, N microphone signals (p. 3 paragraph 31) with a first sampling frequency SF1, each microphone signal comprising a target signal component and a noise signal component (implicit for any microphone signal);

- determining a source location signal (p. 7 paragraph 68);

- deriving a beam shape control signal from a group of consecutive samples of the source location signal (p. 10 paragraph 94, the beamforming task may be distributed to a device; p. 10 paragraph 102);

- generating a beam shaped audio signal from the microphone signals, under control of the beam shape control signal (p. 10 paragraph 94);

- deriving a set of metadata for the bundled audio signal (implicit when transmitting data with a mobile phone);

- generating data packets, wherein a data packets comprises groups of the bundled audio signal and metadata (implicit when transmitting data with a mobile phone, p. 10 paragraph 102);

- wireless streaming of the data packets to the central unit (p. 4 paragraph 36).

2.2 The subject-matter of claim 1 therefore differs from this known speech processing method in that the source location signal is determined with a second sampling frequency SF2, wherein SF1 > SF2.

2.3 The problem to be solved by the present invention may therefore be regarded as providing a signal that requires less bits for storage or transmission.

2.4 The solution proposed in claim 1 of the present application cannot be considered as involving an inventive step. It is common knowledge in the field of audio signal processing and speech processing to downsample a microphone signal to a more compact version with a lower sampling frequency, in order to require less bits for storage or transmission. Therefore, claim 1 is not inventive over a combination of D2 and common knowledge.

3 Lack of inventive step, claim 11

3.1 D2 is considered to be the closest prior art with respect to claim 11, and describes a method in a central unit (server, p. 3 paragraph 31) of a speech processing system comprising P microphone units (multiple devices p. 3 paragraph 31) that perform the method according to any of claims 1-10, the method comprising

- wirelessly receiving P streams with data packets from the P microphone units, each data packet comprising samples of an audio signal and meta data corresponding to the audio (p. 3 paragraph 31, p. 4 paragraph 36, p. 10 paragraph 102 describes beamforming may be distributed to the devices, the meta data is then implicitly disclosed);

- synchronizing the data packets to obtain P synchronized streams of data packets (p. 9 paragraph 93, Fig. 4 step 410);

- detecting in each stream which parts of the bundled audio signals comprise a target signal component of an active speaker that is linked to the microphone unit that generated said stream, based on the audio signals and metadata (p. 10 paragraph 94, Fig. 4 step 415);

- forwarding the detected parts of the audio signals for further processing (Fig. 4 step 420 performing ASR on focused signals).

3.2 The subject-matter of claim 1 therefore differs from this known speech processing method in that D2 does not explicitly describe meta data corresponding to the audio signals.

3.3 The problem to be solved by the present invention may therefore be regarded as transmitting extra information together with the audio signals.

3.4 The solution proposed in claim 11 of the present application cannot be considered as involving an inventive step. It is common knowledge in the field of transmitting audio or audiovisual information to provide additional metadata, such as the commonly known MPEG standard. Furthermore, since the devices in D1 may perform beamforming prior to tranmission, such information is also implicitly transmitted to the central unit. Therefore, claim 1 is not inventive over a combination of D2 and common knowledge.

4 Dependent claims

4.1 Claims 2, 8, 10: D1 implicitly describes receiving the speech profile from the microphone via the central unit, since the processing tasks may include providing a user identifier and focused channel (p. 10 paragraphs 94-95)

4.2 Claims 3, 12: It is a normal design procedure for the metadata and audio signal samples to be recorded in an ith and i + T data package.

4.3 Claim 4: D1 describes directionally focussing on a speaker, which implies that the direction from which the target signal component is coming from is determined in D1.

4.4 Claims 5-7, 9: It is a normal design procedure to start or stop streaming based on whether the required information is present in the data to be streamed.

4.5 Claims 7, 14: D1 describes determining a speaker speech profile, checking whether the profile matches a speech profile of a reference speaker of a microphone (p. 10 paragraph 95),

4.6 Claim 15: Since the microphone units in D1 may also apply processing, the resulting information from the processing must implicitly be then transmitted to the central unit (p. 10 paragraph 102, p. 7 paragraph 68).

4.7 Claim 16: D1 describes a microphone unit for use in a speech processing system comprising P microphone units (one of multiple devices 110A, 110B, 110C, 110D) and a central unit (server, p. 3 paragraph 31), the microphone unit comprising N microphones (p. 3 paragraph 31), a wireless communication unit (p. 4 paragraph 36), processor and memory for storing instructions.

4.8 Claim 17: D1 describes a central unit (server, p. 3 paragraph 31) for use in a speech processing system comprising P microphone units (one of multiple devices 110A, 110B, 110C, 110D) and a central unit (server, p. 3 paragraph 31), the central unit comprising a wireless communication unit (p. 4 paragraph 36) configured to communicate with each of the microphone units, a processor and memory for storing instructions.

4.9 None of the cited prior art documents describe the further details of claim 4, 5-7, 9, 13.