(12) UK Patent Application (19) GB (11) 2510331 (13) A

(43) Date of A Publication 06.08.2014

(21) Application No: 1223238.5

(22) Date of Filing: 21.12.2012

(71) Applicant(s):
Microsoft Corporation
One Microsoft Way, Redmond,
Washington 98052-7329, United States of America

(72) Inventor(s):
Per Ahgren

(74) Agent and/or Address for Service:
Page White & Farrer
Bedford House, John Street, London, WC1N 2BF,
United Kingdom

(51) INT CL:
G10L 21/0208 (2013.01) ***H04B 3/20*** (2006.01)
H04M 9/08 (2006.01)

(56) Documents Cited:
WO 2006/040734 A1 **US 7068780 B1**
US 6256383 B1

(58) Field of Search:
INT CL **G10L, H04B, H04M**
Other: **WPI, EPODOC.**

(54) Title of the Invention: **Echo suppression**
Abstract Title: **Echo suppression in an audio signal**

(57) An audio signal outputted from *eg.* a speaker 210 on a hands-free phoneis received at *eg.* a microphone 212 and the echo cancelled by modelling the echo path according to a Finite Impulse Response (FIR) model 304 and a second, different model (*eg.* an exponential model) 308. Each model estimates the echo power of a first and second component of the echo 306, 310 respectively (*eg.* early and later echoes, fig.5), then those estimates are combined 312 and used to apply echo suppression 314 to the audio signal. The models may be dynamically adapted based on the output and received audio signals x(t), y(t).

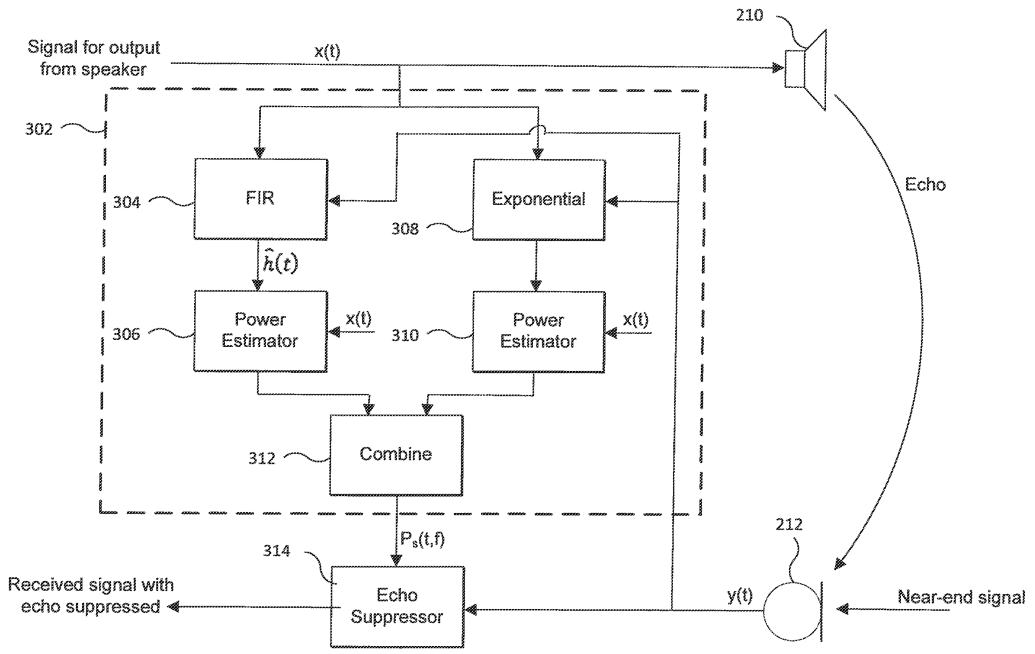

FIG. 3

GB 2510331 A

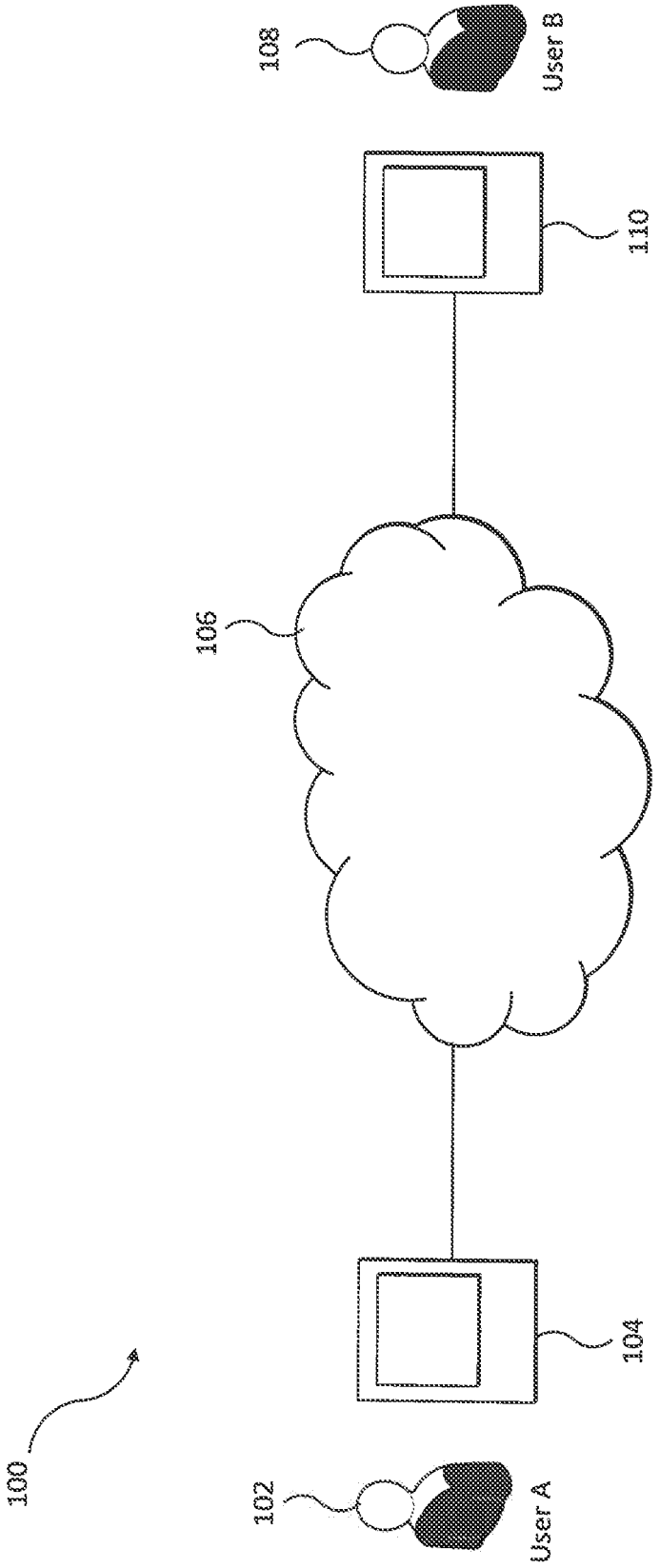

FIG. 1

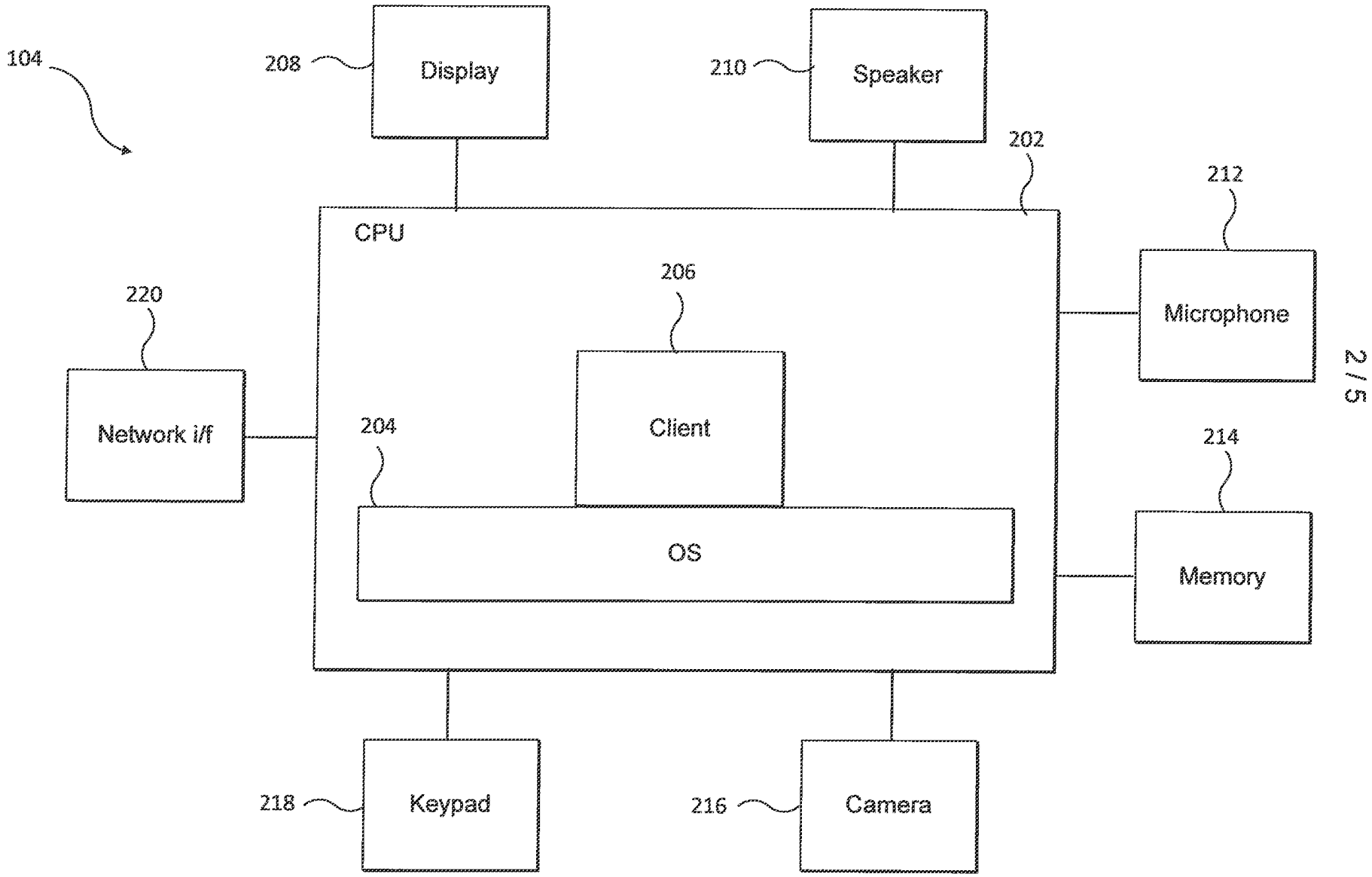
104
208
Display
210
Speaker
202
212
CPU
206
220
Microphone
Network i/f
204
Client
214
OS
Memory
218
Keypad
216
Camera

FIG. 2

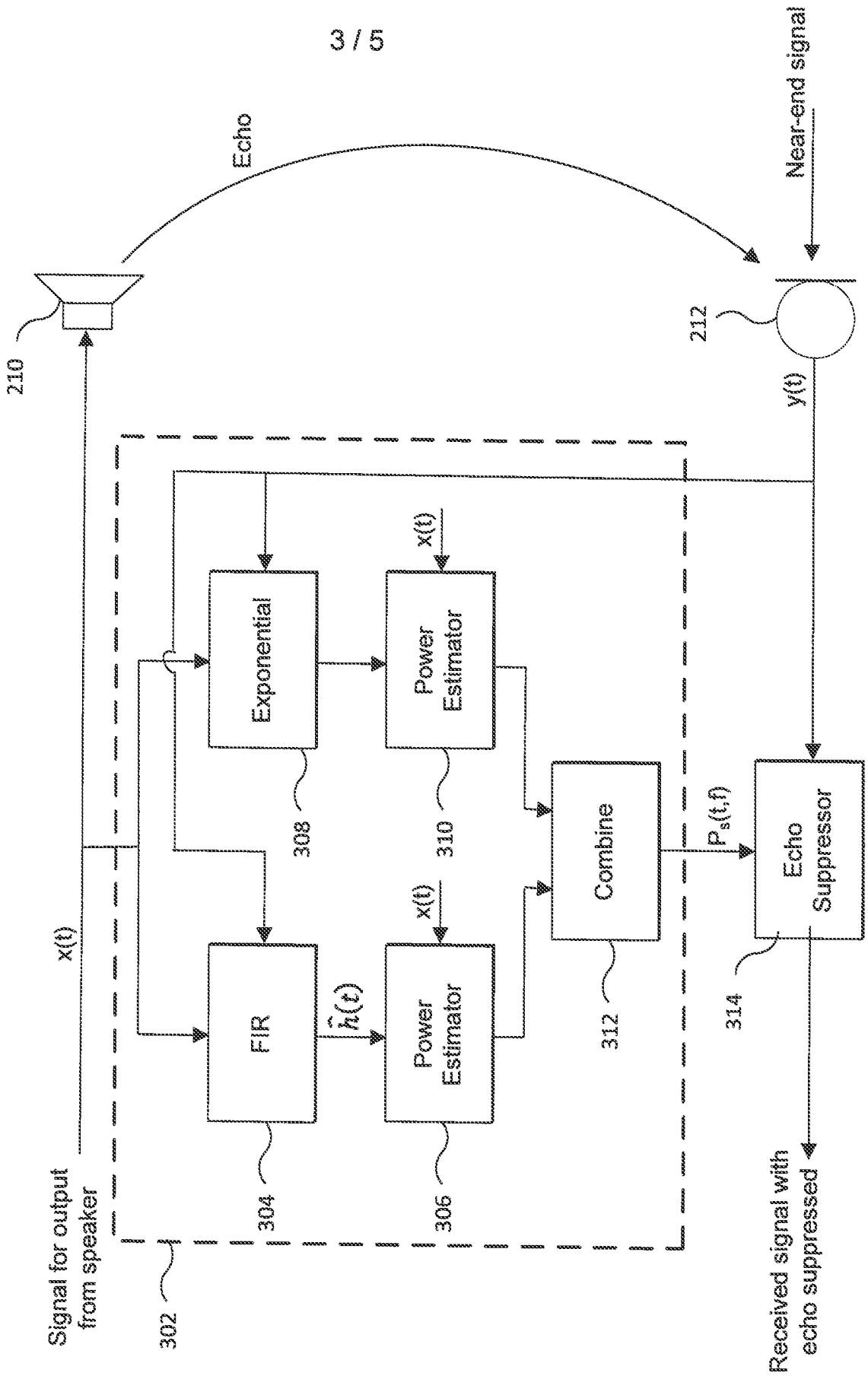
Echo
Near-end signal
210
212
y(t)
x(t)
Exponential
Power Estimator
x(t)
308
310
FIR
$\hat{h}(t)$
Power Estimator
x(t)
Combine
$P_s(t,f)$
Echo Suppressor
312
314
304
306
302
Signal for output from speaker
Received signal with echo suppressed

FIG. 3

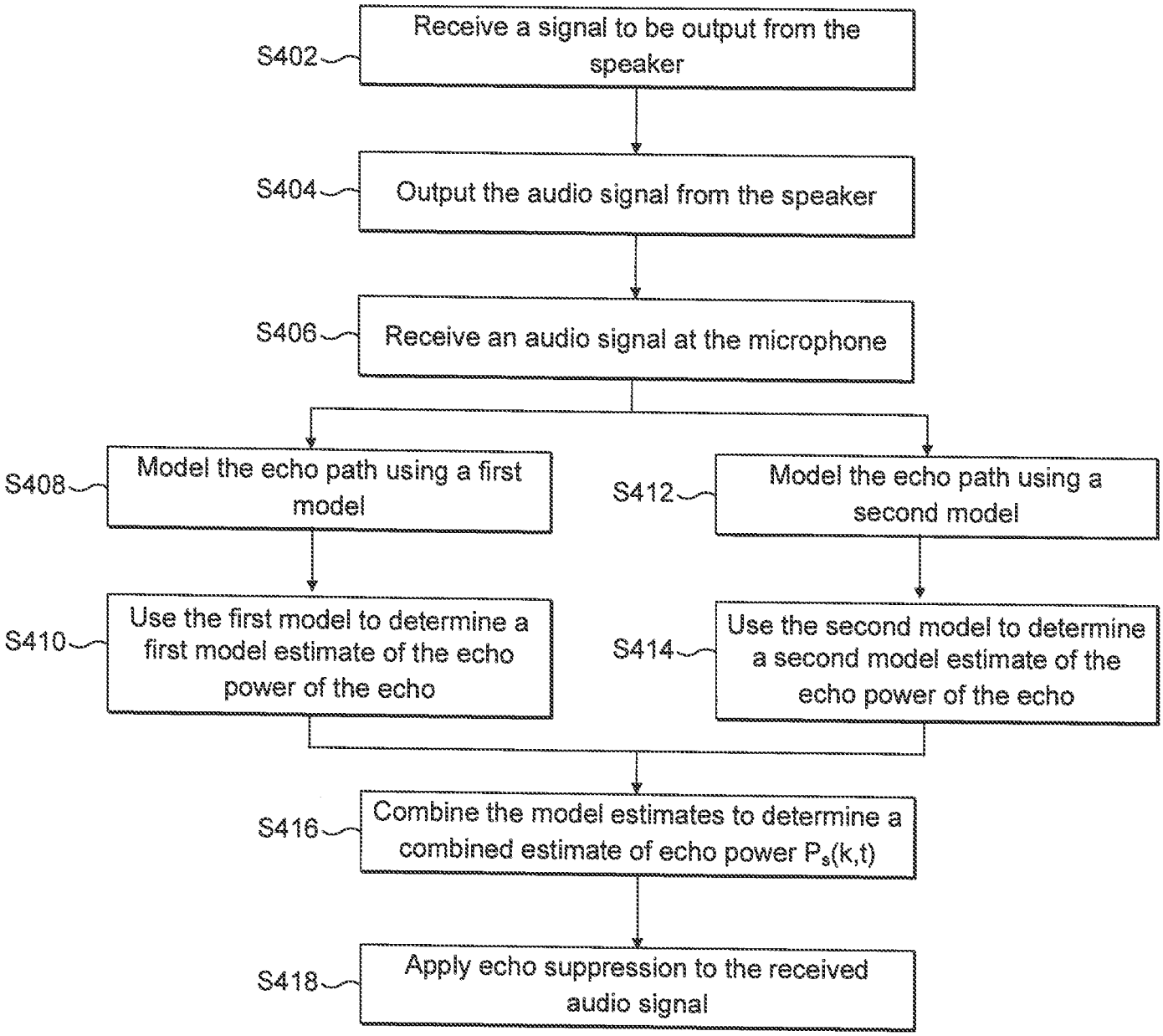
S402
Receive a signal to be output from the speaker
S404
Output the audio signal from the speaker
S406
Receive an audio signal at the microphone
S408
Model the echo path using a first model
S410
Use the first model to determine a first model estimate of the echo power of the echo
S412
Model the echo path using a second model
S414
Use the second model to determine a second model estimate of the echo power of the echo
S416
Combine the model estimates to determine a combined estimate of echo power $P_s(k,t)$
S418
Apply echo suppression to the received audio signal

FIG. 4

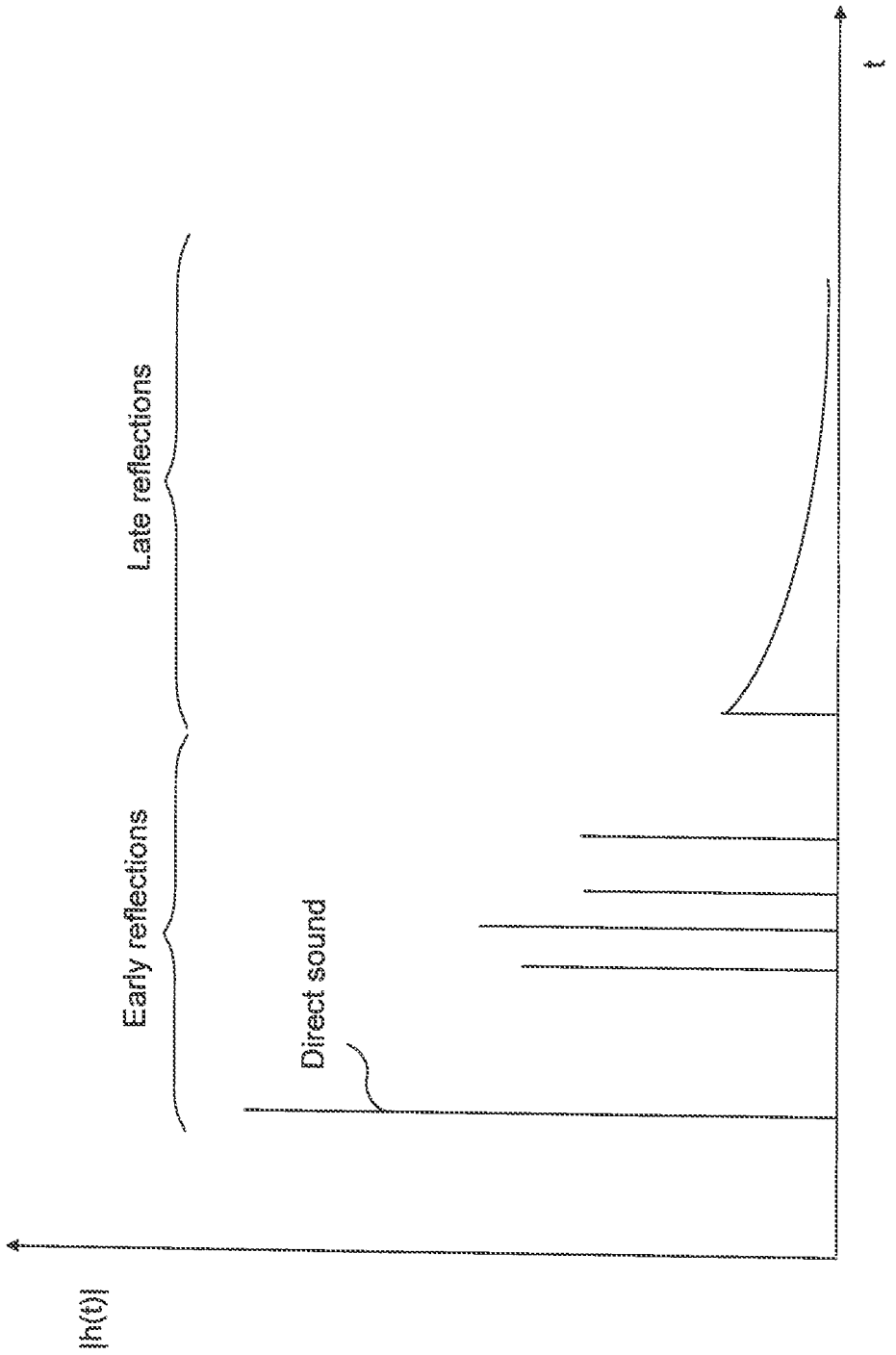

FIG. 5

ECHO SUPPRESSION

Background

A device may have audio input apparatus that can be used to receive audio signals from the surrounding environment. The device may also have audio output apparatus that can be used to output audio signals to the surrounding environment. For example, a device may have one or more speakers for outputting audio signals and one or more microphones for receiving audio signals. Audio signals which are output from the speaker(s) of the device may be received as "echo" in the audio signal received by the microphone(s). It may be the case that this echo is not desired in the received audio signal. For example, the device may be a user device (such as a mobile phone, tablet, laptop, PC, etc) which is used in a communication event, such as an audio or video call, with another user device over a network. Far-end signals of the call may be output from the speaker at the user device and may be received as echo in the audio signals received by the microphone at the device. Such echo can be disturbing to users of the call, and the perceived quality of the call may be reduced due to the echo. In particular, the echo may cause interference for near-end audio signals which are intended to be received by the microphone and transmitted to the far-end in the call. Therefore echo cancellation and/or echo suppression may be applied to the received audio signals to thereby suppress the echo in the received audio signal. The power of the echo in the received audio signal may vary depending upon the arrangement of the user device. For example, the user device may be a mobile phone and in that case, the power of the echo in the received audio signal would normally be higher when the mobile phone is operating in a "hands-free" mode compared to when the mobile phone is not operating in a "hands-free" mode.

Echo cancellation (or "echo subtraction") techniques aim to estimate an echo signal included in the audio signal received at the microphone, based on knowledge of the audio signal which is output from the speaker. The estimate of the echo signal can then be subtracted from the received audio signal thereby removing at least some of the echo from the received audio signal.

Echo suppression is used to apply frequency-dependent suppression to the received audio signal to thereby suppress the echo in the received audio signal. In order for echo suppression to be implemented effectively, an echo suppressor needs to have an accurate estimate of the power of the echo in the received audio signal.

Summary

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.

There is provided a method of suppressing echo in a received audio signal. As part of the echo suppression, an echo path of the echo is modelled using two different models, wherein a first of the models is a Finite Impulse Response (FIR) based model. The first model is used to determine a first model estimate of the echo power of at least a first component of the echo in the received audio signal. The second model is used to determine a second model estimate of the echo power of at least a second component of the echo in the received audio signal. The first and second model estimates of the echo power are combined to determine a combined estimate of the echo power of the echo. The combined estimate of the echo power of the echo is used to apply echo suppression to the received audio signal, thereby suppressing the echo in the received audio signal.

The method may be used in a call (e.g. a call implementing voice over internet protocol (VoIP) to transmit audio data between user devices) in which case the outputted audio signal may be a far-end signal received from the far-end of the call, and the received signal includes the resulting echo and a near-end signal for transmission to the far-end of the call.

Brief Description of the Drawings

For a better understanding of the present invention and to show how the same may be put into effect, reference will now be made, by way of example, to the following drawings in which:

Figure 1 shows a schematic illustration of a communication system;
Figure 2 is a schematic block diagram of a user device;
Figure 3 is a functional diagram showing modules of a user device for use in echo suppression;
Figure 4 is a flow chart for a process of suppressing echo; and
Figure 5 is a graph representing an impulse response of an echo signal.

Detailed Description of Preferred Embodiments

Preferred embodiments of the invention will now be described by way of example only.

In order for echo suppression to be implemented effectively, an echo suppressor needs to have an accurate estimate of the power of the echo in the received audio signal. One way to estimate the echo power is to apply an FIR filter, either in the time-domain or in the power frequency domain. This filter could in turn either be adapted in the time-domain or in the power frequency domain. For long echo paths this has the drawback that the FIR filter adaptation and the corresponding echo power estimation becomes computationally complex for long echo paths because the complexity of the FIR filter adaptation and corresponding echo power estimation is proportional to the length of the echo path for the echo that is to have its power content estimated.

Therefore, in accordance with methods described herein a plurality of models (e.g. two models) are used for modelling the echo path. The different models may model different components of the echo in the received audio signal. This allows each model to be chosen to suit different types of echo component in the echo. For example, a first model may model an early reflections component of the echo in the received audio signal, whilst a second model may model a late

reflections component of the echo in the received audio signal. Figure 5 is a graph representing the magnitude of an impulse response of an echo signal $|h(t)|$. It can be seen that there is a strong response corresponding to the direct path between a speaker and a microphone, and there are a number of strong responses corresponding to early reflections of the echo signal (e.g. a single or double reflection off surfaces such as walls) between the speaker and the microphone. As shown in Figure 5, the early reflections component of the echo corresponds to the direct path and the first few reflections of the echoes against surfaces. It can also be seen in Figure 5 that the late reflections tend to blur together resulting in reverberation rather than a number of distinct, distinguishable reflections. Since the different components of the echo have different characteristics (e.g. early reflections have distinct peaks whereas late reflections blur together in reverberation), different models of the echo path may be used to model the different components of the echo. In this way the different models can be chosen to suit the characteristics of a particular component of the echo which will be modelled by the model. The late reflections can be modelled well with an exponential model, whereas the early reflections tend not to be well modelled with an exponential model. In one example, there are two models: (i) a FIR based model that is used to model the early reflections part of the echo path, and (ii) an exponential model that is used to model the late reflections part of the echo path. The FIR based model may be more accurate than the exponential model in modelling the echo path of the echo (in particular for the early reflections component of the echo). However, as the length of the echo path increases, the complexity of the FIR model increases more than the complexity of the exponential model. Since the exponential model has a very low computational complexity, it is suited for use in modelling very long echo paths at a very low computational complexity.

As described herein, the power of at least the early reflections component of the echo can be estimated using the output from an FIR filter that is adapted to approximate the impulse response of the echo path between a loudspeaker outputting audio signals and a microphone receiving audio signals including the echo resulting from the outputted audio signals.

The FIR filter may be used to estimate the echo signal, from which the echo power can be estimated and used in an echo suppression method. However, the FIR filter might only be used to estimate the echo power, and not to estimate the actual echo signal. This may be advantageous because the requirements for accuracy in the FIR filter are much less when used to estimate the echo power compared to if the FIR filter is used to estimate the actual echo signal. Therefore by estimating the echo power (rather than the echo signal) from the FIR filter, echo suppression is more robust to problems such as clock-drift between the playout (e.g. from a loudspeaker) and recording sides (e.g. at a microphone) in the VoIP client, nonlinearities in the echo path and changes in the echo path. In embodiments described herein, the FIR filter is adapted using time-domain data including phase information.

Figure 1 shows a communication system 100 comprising a first user 102 ("User A") who is associated with a first user device 104 and a second user 108 ("User B") who is associated with a second user device 110. In other embodiments the communication system 100 may comprise any number of users and associated user devices. The user devices 104 and 110 can communicate over the network 106 in the communication system 100, thereby allowing the users 102 and 108 to communicate with each other over the network 106. The communication system 100 shown in Figure 1 is a packet-based communication system, but other types of communication system could be used. The network 106 may, for example, be the Internet. Each of the user devices 104 and 110 may be, for example, a mobile phone, a tablet, a laptop, a personal computer ("PC") (including, for example, Windows™, Mac OS™ and Linux™ PCs), a gaming device, a television, a personal digital assistant ("PDA") or other embedded device able to connect to the network 106. The user device 104 is arranged to receive information from and output information to the user 102 of the user device 104. The user device 104 comprises output means such as a display and speakers. The user device 104 also comprises input means such as a keypad, a touch-screen, a microphone for receiving audio signals and/or a camera for capturing images of a video signal. The user device 104 is connected to the network 106.

The user device 104 executes an instance of a communication client, provided by a software provider associated with the communication system 100. The communication client is a software program executed on a local processor in the user device 104. The client performs the processing required at the user device 104 in order for the user device 104 to transmit and receive data over the communication system 100.

The user device 110 corresponds to the user device 104 and executes, on a local processor, a communication client which corresponds to the communication client executed at the user device 104. The client at the user device 110 performs the processing required to allow the user 108 to communicate over the network 106 in the same way that the client at the user device 104 performs the processing required to allow the user 102 to communicate over the network 106. The user devices 104 and 110 are endpoints in the communication system 100. Figure 1 shows only two users (102 and 108) and two user devices (104 and 110) for clarity, but many more users and user devices may be included in the communication system 100, and may communicate over the communication system 100 using respective communication clients executed on the respective user devices.

Figure 2 illustrates a detailed view of the user device 104 on which is executed a communication client instance 206 for communicating over the communication system 100. The user device 104 comprises a central processing unit ("CPU") or "processing module" 202, to which is connected: output devices such as a display 208, which may be implemented as a touch-screen, and a speaker (or "loudspeaker") 210 for outputting audio signals; input devices such as a microphone 212 for receiving audio signals, a camera 216 for receiving image data, and a keypad 218; a memory 214 for storing data; and a network interface 220 such as a modem for communication with the network 106. The user device 104 may comprise other elements than those shown in Figure 2. The display 208, speaker 210, microphone 212, memory 214, camera 216, keypad 218 and network interface 220 may be integrated into the user device 104 as shown in Figure 2. In alternative user devices one or more of the display 208, speaker 210, microphone 212, memory 214, camera 216, keypad

218 and network interface 220 may not be integrated into the user device 104 and may be connected to the CPU 202 via respective interfaces. One example of such an interface is a USB interface. If the connection of the user device 104 to the network 106 via the network interface 220 is a wireless connection then the network interface 220 may include an antenna for wirelessly transmitting signals to the network 106 and wirelessly receiving signals from the network 106.

Figure 2 also illustrates an operating system ("OS") 204 executed on the CPU 202. Running on top of the OS 204 is the software of the client instance 206 of the communication system 100. The operating system 204 manages the hardware resources of the computer and handles data being transmitted to and from the network 106 via the network interface 220. The client 206 communicates with the operating system 204 and manages the connections over the communication system. The client 206 has a client user interface which is used to present information to the user 102 and to receive information from the user 104. In this way, the client 206 performs the processing required to allow the user 102 to communicate over the communication system 100.

With reference to Figures 3 and 4 there is now described a method of suppressing echo. Figure 3 is a functional diagram of a part of the user device 104 showing how an echo suppression process is implemented, and Figure 4 is a flow chart for the process of suppressing echo.

As shown in Figure 3, the user device 104 comprises the speaker 210, the microphone 212, a modelling module 302 and an echo suppression module 314. The modelling module 302 comprises a FIR filter module 304, a first power estimating module 306, an exponential filter module 308, a second power estimating module 310 and a combining module 312. A signal x(t) to be output from the speaker 210 is coupled to an input of the speaker 210. It should be noted that in the embodiments described herein there is just one speaker (indicated by reference numeral 210 in the figures) but in other embodiments there may be more than one speaker to which the signal to be outputted is coupled (for outputting therefrom). Similarly, in the embodiments described

herein there is just one microphone (indicated by reference numeral 212 in the figures) but in other embodiments there may be more than one microphone which receive audio signals from the surrounding environment. The signal to be output from the speaker 210 is also coupled to the modelling module 302. In particular, the signal to be output from the speaker 210 is coupled to a first input of the FIR filter module 304 and to a first input of the exponential filter module 308. An output of the microphone 212 is coupled to the modelling module 302. In particular, the output of the microphone 212 is coupled to a second input of the FIR filter module 304 and to a second input of the exponential filter module 308 to a first input of the first power estimating module 306 and to a first input of the second power estimating module 310. The output of the microphone 212 is also coupled to a first input of the echo suppression module 314. An output of the FIR filter module 304 is coupled to a second input of the first power estimating module 306. An output of the exponential filter module 308 is coupled to a second input of the second power estimating module 310. An output of the first power estimating module 306 is coupled to a first input of the combining module 312. An output of the second power estimating module 310 is coupled to a second input of the combining module 312. An output of the modelling module 302 is coupled to a second input of the echo suppression module 314. In particular an output of the combining module 312 is coupled to the second input of the echo suppression module 314. An output of the echo suppression module 314 is used to provide the received signal (with echo suppression having been applied) for further processing in the user device 104.

In step S402 a signal is received which is to be outputted from the speaker 210. For example, the signal to be outputted may be a far-end signal that has been received at the user device 104 from the user device 110 during a call between the users 102 and 108 over the communication system 100. Any processing that is required to be performed on the received signal (e.g. decoding using a speech codec, depacketizing, etc) is performed as is known in the art (e.g. by the client 206) to arrive at the signal x(t) which is suitable to be outputted from the speaker 210. The signal x(t) is a digital signal. At least some of the processing of the signal in the user device 104 prior to outputting the signal from the speaker 210 is performed in the digital domain. As is known in the art,

a digital to analogue converter (DAC) is applied to the digital signal x(t) before playout from the loudspeaker 210. Similarly, an analogue to digital converter (ADC) is applied to the signal captured by the microphone 212 to arrive at the digital signal y(t).

In other embodiments, the signal to be outputted may be received from somewhere other than over the communication system 100 in a call. For example, the signal to be outputted may have been stored in the memory 214 and step S402 may comprise retrieving the signal from the memory 214.

In step S404 the audio signal x(t) is outputted from the speaker 210. In this way the audio signal x(t) is outputted to the user 102.

In step S406 the microphone 212 receives an audio signal. As shown in Figure 3 the received audio signal may include a near-end signal which is a desired signal or "primary signal". The near-end signal is the signal that the user 102 intends the microphone 212 to receive. However, the received audio signal also includes an echo signal resulting from the audio signals outputted from the speaker 210 in step S404. The received audio signal may also include noise, such as background noise. Therefore, the total received audio signal y(t) can be given by the sum of the near-end signal, the echo and the noise. The echo and the noise act as interference for the near-end signal.

The FIR filter module 304 takes as inputs the outputted audio signal x(t) and the received audio signal y(t). In step S408 the FIR filter module 304 dynamically adapts a FIR filter estimate $\hat{h}(t)$ in the time domain based on the outputted audio signal x(t) and the received audio signal y(t) to model an echo path $h(t)$ of the echo in the received audio signal y(t). The "impulse response of the echo path h(t)" is also referred to herein as the "echo path h(t)". The FIR filter module 304 is used to model the early reflections component of the echo in the received audio signal y(t). In order to do this, the length of the FIR filter used by the FIR filter module 304 to model the echo path has a finite length L which is long enough to model the early reflections component of the echo, but not to fully model the late reflections component of the echo (see Figure 5). In this

way the length of the FIR model used by the FIR filter module 304 does not need to be as long as the full echo path of the echo in the received audio signal y(t). This may ensure that the complexity of the FIR model does not become too large.

For an approximately linear echo path, the echo path h(t) describes how the echo in the received audio signal relates to the audio signal x(t) output from the speaker 210, e.g. according to the equation $y^{echo}(t) = \sum_{n=0}^{N_{true}} h_n(t)x(t-n)$, where $y^{echo}(t)$ is the echo in the received audio signal y(t), N_{true} is the number of samples of the outputted signal x(t) which are received by the microphone 212 and $h_n(t)$ are weights describing the echo path h(t). The echo path h(t) may vary in both time and frequency and may be referred to herein as h(t) or h(t,f). The echo path h(t) may depend upon (i) the current environmental conditions surrounding the speaker 210 and the microphone 212 (e.g. whether there are any physical obstructions to the passage of the audio signal from the speaker 210 to the microphone 212, the air pressure, temperature, wind, etc), and (ii) characteristics of the speaker 210 and/or the microphone 212 which may alter the signal as it is outputted and/or received.

The FIR filter module 302 models the early reflections component of the echo path h(t) of the echo in the received audio signal by determining a weighted sum of the current and a finite number (N) of previous values of the outputted audio signal x(t). The FIR filter module 302 therefore implements an Nth order FIR filter which has a finite length (in time) over which it considers the values of the outputted audio signal x(t) in determining the estimate of the early reflections component of the echo path $\hat{h}(t)$. In this way, the FIR filter module 302 dynamically adapts the FIR filter estimate $\hat{h}(t)$. The operation is described by the following equation, which defines the echo in the received audio signal y(t) in terms of the outputted audio signal x(t):

$$\hat{y}^{echo}(t) = \sum_{n=0}^{N} \hat{h}_n(t)x(t-n).$$

Therefore N+1 samples of the outputted audio signal x(t) are used, with a respective N+1 weights $\hat{h}_n(t)$. The set of N+1 weights $\hat{h}_n(t)$ is referred to herein simply as the estimate of the echo path $\hat{h}(t)$. In other words the estimate of the echo path $\hat{h}(t)$ is a vector having N+1 values where the FIR filter module

302 implements an Nth order FIR filter, taking N+1 values (e.g. N+1 frames) of the signal x(t) into account.

It can be appreciated that it is easier to adapt the FIR filter estimate $\hat{h}(t)$ when the echo is a dominant part of the received audio signal, that is when $y(t) \cong y^{echo}(t)$. For example, in some embodiments it may be possible to detect when the power of the near-end signal is greater than the power of the echo (e.g. when the user 102 is speaking), and whilst that is the case the FIR estimate $\hat{h}(t)$ is not adapted, but when the power of the near-end signal is less than the power of the echo in the received audio signal y(t) (e.g. when the user 102 is not speaking) the FIR estimate $\hat{h}(t)$ is adapted.

However, it may be possible to adapt the FIR filter estimate $\hat{h}(t)$ even when the echo is not a dominant part of the received audio signal.

The FIR filter estimate $\hat{h}(t)$ is passed from the FIR filter module 304 to the first power estimating module 306. The first power estimating module 306 estimates the echo power of the early reflections component of the echo in the received audio signal in one of at least two ways, as described below.

In one method, in step S410 the power estimating module 304 estimates the echo power of the early reflections component of the echo in the received audio signal based on the filter estimate $\hat{h}(t)$ determined in step S408 and based on the input signal x(t). Step S410 might not comprise estimating the echo signal $y^{echo}(t)$ in the received audio signal y(t). The echo power of the early reflections component of the echo is estimated as a function of time and frequency. In echo suppression a rather low accuracy of the echo power estimate is sufficient to achieve good echo suppression. According to methods described herein the power response can be computed in a way that is less sensitive to problems. Furthermore, the power response can be estimated in a different way than the actual echo path would be estimated. For example, the power response for a frequency, f, may be computed using the estimate of the FIR filter for that frequency ,f. Alternatively or additionally, the estimate of the FIR filter for the frequency ,f, may be used to compute the power response for a

different frequency, v, where $v \neq f$. In other words, the method may include using an extrapolated echo path power response that is computed for another frequency region than the one where it is applied. In this sense, the power response is computed based on the FIR filter estimate, although some extrapolation may be required to determine the power response for a particular frequency. That is to say, an FIR filter estimate obtained for a certain frequency region may be used to compute a predicted (or extrapolated) power response estimate for another frequency region, .i.e., the power response used to estimate the echo power is not necessary the power response of the FIR filter but could also be a power response (e.g. for a different frequency region) that is computed based on the FIR filter.

Step S410 may comprise estimating the echo power of the early reflections component of the echo $P_s^{early}(t,f)$, which is a scalar power bin with values for time t and frequency f, and which is calculated for a particular frequency and time according to the equation:

$$\hat{P}_s^{early}(t,f) = \left(\hat{y}_{echo}(t,f)\right)^2 = \left(\sum_{n=0}^{N} \hat{h}_n(t,f)x(t-n,f)\right)^2,,$$

In this way, the FIR filter estimate $\hat{h}(t)$, that has been adapted using the speaker and microphone signals x(t) and y(t) to approximate the time-varying echo path h(t) of the VoIP client, is used, with the outputted audio signal samples x(t), to estimate the power $P_s^{early}(t,f)$ of the early reflections component of the echo signal at time t and frequency f.

In a second method, in step S410 the first power estimating module 306 determines at least one power response from the FIR filter estimate $\hat{h}(t)$. The power response information is determined by analysing the FIR filter estimate $\hat{h}(t)$. The power response (or "frequency response") gives an indication of the power response of the echo path h(t) as a function of frequency. Note that although the echo path is denoted h(t) herein, this is for simplicity, and it is re-iterated that the echo path h(t) and the estimate of the echo path $\hat{h}(t)$ are functions of both time and frequency.

Then further in step S410 the first power estimating module 306 estimates the echo power of the early reflections component of the echo in the received audio signal based on the determined power response(s). Indeed, step S410 might not comprise estimating the echo signal $y^{echo}(t)$ in the received audio signal y(t). The echo power of the early reflections component of the echo is estimated as a function of time and frequency.

In particular, the FIR filter estimate $\hat{h}(t)$ has a length L in the time domain. Step S410 comprises partitioning the FIR filter estimate $\hat{h}(t)$ into a plurality (P) of partitions in the time domain of length L/P each. Each of the partitions of the FIR filter estimate $\hat{h}(t)$ is transformed into the frequency domain and squared to determine a respective power response $|\hat{H}_p(f)|^2$ in the frequency domain for each of the partitions. It can therefore be appreciated that $|\hat{H}_p(f)|^2$ is the frequency response of partition p.

Step S410 comprises estimating the echo power of the echo in the received audio signal by performing a weighted sum of a plurality of measures of the power of a respective plurality of frames of the outputted audio signal, wherein the weights in the sum are given by respective ones of the power responses $|\hat{H}_p(f)|^2$.

Therefore, the estimate of the echo power $P_s^{early}(k, f)$ of the early reflections component of the echo in the received audio signal, for a frame k, can be estimated in step S410 according to the equation:

$$\hat{P}_s^{early}(k, f) = \sum_{p=0}^{P-1} |\hat{H}_p(f)|^2 |X(k-p, f)|^2,$$

where $|X(k-p, f)|^2$ is the power spectral density of the loudspeaker signal for frame k – p. The frame index k is a measure of the time, and as such $P_s^{early}(k, f)$ can be rewritten to be a function of time rather than of frame indices, to give the estimate $\hat{P}_s^{early}(t, f)$ of the echo power, and vice versa.Note that in order for the above equation to be correct, the length of the filter partitions and loudspeaker signals used to compute $|\hat{H}_p(f)|^2$ and $|X(k-p, f)|^2$ should be carefully selected in order to minimise circular convolution effects. This

selection is performed to ensure that the lengths of the partition of h and the loudspeaker input signal frame are properly matched to the length of the microphone signal used to adapt the filter estimate h, and for which the echo power is to be estimated. If each partition is of length P, each loudspeaker (X) frame length is M, and each microphone signal frame length is N, the typical requirement to being able to avoid circular convolution effects is that N+P-1<M. Although this selection is preferable, the methods described herein will work regardless of the selection because circular convolution effects are ignored in the methods.

In this way, the FIR filter estimate $\hat{h}(t)$, that has been adapted using the speaker and microphone signals x(t) and y(t) to approximate the time-varying echo path h(t) of the VoIP client, is used with the outputted audio signals x(t) to determine the power responses $|\hat{R}_p(f)|^2$ which are then used to estimate the power $P_s^{early}(t,f)$ of the early reflections component of the echo signal at time t and frequency f.

In general a linear (e.g., FIR) model is used to model the early reflections part of the echo path which is then used to model the echo power corresponding to the early reflections. There are many ways to do this and the scope of this disclosure is not limited to the examples given above. The echo power estimate from the early reflections can be estimated as a function of time and frequency.

When the first power estimating module 306 has determined it according to any of the methods described above, the estimate $\hat{P}_s^{early}(t,f)$ of the echo power of the early reflections component of the echo is output from the first power estimating module 306 and received by the combining module 312.

The exponential filter module 308 takes as inputs the outputted audio signal x(t) and the received audio signal y(t). In step S412 the exponential filter module 308 is used to model at least the late reflections component of the echo in the received audio signal y(t). As can be seen in Figure 5, the late reflections component of the echo (which comprises mainly reverberation) is well suited to being modelled with an exponential model.

Step S412 comprises determining an estimate of a decay factor $\hat{\gamma}(f)$ of the exponential model based on the outputted audio signal x(t) and the echo in the received audio signal $y^{echo}(t)$. The estimated decay factor $\hat{\gamma}(f)$ is a function of frequency but it may also vary in time. The estimate of the decay factor is adapted dynamically. Similarly to as described above in relation to the adaptation of the FIR model in step S408, it is easier to adapt the estimate of the decay factor when the echo is a dominant part of the received audio signal, that is when $y(t) \cong y^{echo}(t)$.

The decay factor $\hat{\gamma}(f)$ can be estimated using some estimation method (an example would be by fitting a linear line to the logarithm of the tail of the power response of the FIR filter constituting the model for the early reflections component of the echo. It will be apparent that other techniques could be used to determine the estimate of the decay factor $\hat{\gamma}(f)$.

The estimate of the decay factor of the exponential model $\hat{\gamma}(f)$ is passed from the exponential filter module 308 to the second power estimating module 310. In step S414 the second power estimating module 310 uses the estimate of the decay factor of the exponential model $\hat{\gamma}(f)$ to determine a second model estimate of the echo power of the echo. In particular, the second power estimating module 310 uses the estimate of the decay factor of the exponential model $\hat{\gamma}(f)$ to determine the echo power $P_s^{late}(k, f)$ of the late reflections component of the echo in the received audio signal y(t). k is the frame index, and therefore gives an indication of time.

For example, the second power estimating module 310 may estimate the power $P_s^{late}(k, f)$ of the late components of the echo according to the equation:

$$\hat{P}_s^{late}(k, f) = \sum_{i=1}^{\infty} \hat{\gamma}^i(f) \hat{P}_{s,end}^{early}(k - i, f) = \hat{\gamma}(f) \hat{P}_{s,end}^{early}(k - 1, f) + \hat{\gamma}(f) \hat{P}_s^{late}(k - 1, f)$$

where $\hat{P}_{s,end}^{early}(k, f)$ is the estimate of the power contribution of the last part of the model for the early reflections. $\hat{\gamma}(f)$ is the estimated decay factor for the exponential decay model at a particular time (i.e. for a particular frame). $\hat{\gamma}(f)$ is

an exponential decay and not a weighting function. The equation determines the recursion of the exponentially decaying power. For the equation above to be precisely correct the decay factor $\hat{\gamma}(f)$ must be constant. The superscript *i* in $\hat{\gamma}^i(f)$ means that $\hat{\gamma}(f)$ is raised to the power *i*. In practice, $\hat{\gamma}(f)$ may vary over time. In that case the equation above can be used assuming that $\hat{\gamma}(f)$ is only slowly varying in time.

As described above, $\hat{P}_{s,end}^{early}(k,f)$ is the estimate of the power contribution of the "last part" of the model for the early reflections. In other words, $\hat{P}_{s,end}^{early}(k,f)$ is an estimate of the echo power of a last part of the estimate of the power of the early reflections of the echo in the received audio signal determined using the FIR based model for frame $k-1$, the last part corresponding to the latest of the reflections in the early reflections component of the echo. In this way the exponential model uses the latest of the early reflections component of the echo (e.g. at the boundary of the early reflections and late reflections components shown in Figure 5) and then assumes that this echo component exponentially decays over time in order to determine the late reflections component of the echo.

For example, in the case described above in which the estimate of the power of the early reflections component of the echo in the received audio signal is determined according to the equation $\hat{P}_s^{early}(k,f) = \sum_{p=0}^{P-1}|\hat{H}_p(f)|^2 |X(k-p,f)|^2$, then the estimate $\hat{P}_{s,end}^{early}(t,f)$ of the power contribution of the last part of the model for the early reflections (at time t corresponding to frame k) can be computed as:

$$\hat{P}_{s,end}^{early}(t,f) = |\hat{H}_{P-1}(f)|^2 |X(k-P+1,f)|^2,$$

where $|\hat{H}_{P-1}(f)|^2$ is the power response for the last partition (P-1) considered by the FIR filter model and $|X(k-P+1,f)|^2$ is the measure of the power of the frame which is P-1 before the current frame k of the outputted audio signal. Looking back P-1 frames before the current frame k is as far back in time as the FIR filter considers. To consider echo having a longer echo path than this the

exponential model is used whereby the value of the echo for the frame which is P-1 before the current frame k is assumed to exponentially decay.

When the second power estimating module 310 has determined it according to any of the methods described above, the estimate $\hat{P}_s^{late}(k,f)$ of the echo power of the late reflections component of the echo is output from the second power estimating module 310 and received by the combining module 312.

In step S416 the combining module 312 combines the estimate $\hat{P}_s^{early}(k,f)$ of the echo power of the early reflections component of the echo and the estimate $\hat{P}_s^{late}(k,f)$ of the echo power of the late reflections component of the echo to determine a combined estimate $\hat{P}_s(k,f)$ of the echo power. The combined estimate $\hat{P}_s(k,f)$ of the echo power gives an estimate of the echo power of the echo in the received audio signal taking account of the early and late reflections components (which may constitute all of the components) of the echo. That is, the combined estimate of the echo power gives an estimate of the total echo in the received audio signal. The combination of the estimate $\hat{P}_s^{early}(k,f)$ of the echo power of the early reflections component of the echo and the estimate $\hat{P}_s^{late}(k,f)$ of the echo power of the late reflections component of the echo may be performed as a sum of these two estimates. That is, the combined estimate $\hat{P}_s(k,f)$ of the echo power can be determined according to the equation:

$$\hat{P}_s(k,f) = \hat{P}_s^{early}(k,f) + \hat{P}_s^{late}(k,f).$$

The combined estimate $\hat{P}_s(k,f)$ of the echo power of the echo in the received audio signal is passed from the combining module 312 to the echo suppression module 314. The echo suppression module 314 also receives the audio signal y(t) from the microphone 212. In step S418 the echo suppression module 314 uses the estimate $\hat{P}_s(k,f)$ of the echo power to apply echo suppression to the received audio signal y(t), thereby suppressing the echo in the received audio signal. The estimate $\hat{P}_s(k,f)$ of the echo power is frequency dependent and the suppression applied by the echo suppression module 306 is also frequency dependent. As described above, the frame index k is a measure of time and as such $\hat{P}_s(k,f)$ can be re-written in terms of time to give $\hat{P}_s(t,f)$.

The purpose of the echo suppressor is to suppress the loudspeaker echo present in the microphone signal, e.g. in a VoIP client, to a level sufficiently low for it not to be noticeable/disturbing in the presence of the near-end sounds (non-echo sounds) picked up by the microphone 212. In order to be able to choose the proper amount of echo suppression a good estimate of the echo power (e.g. as a function of frequency and time) is needed, and as described above this is provided to the echo suppression module 314 by the power combining module 312. The echo suppression module 314 is designed to apply signal dependent suppression that varies both over time and frequency to the received audio signal y(t). Echo suppression methods are known in the art. Furthermore, the echo suppression method applied by the echo suppression module 314 may be implemented in different ways. As such, the exact details of the echo suppression method are therefore not described in detail herein.

The echo suppression module 314 outputs the received signal, with the echo having been suppressed, for further processing at the user device 104. For example, the signal output from the echo suppression module 314 may be processed by the client 206 (e.g. encoded and packetized) and then transmitted over the network 106 to the user device 110 in a call between the users 102 and 108. Additionally or alternatively, the signal output from the echo suppression module 314 may be used for other purposes by the user device 104, e.g. the signal may be stored in the memory 214 or used as an input to an application which is executing at the user device 104.

There is therefore described herein the use of two separate models (e.g. an FIR filter module 304 and an exponential filter module 308) to model the echo path to estimate the power of the loudspeaker echo signal in frequency bands picked up by the microphone 212, for the purpose of computing and applying an echo suppression effect/filter (e.g. for use by the VoIP client 206). In examples described herein, a hybrid model consisting of one FIR based model, and one exponential model, is used to estimate the echo power of the echo in the received audio signal. The FIR model is used to model the early reflections component of the echo that corresponds to the first few reflections of the

echoes against surfaces, and that cannot be well approximated to be exponentially decaying. The exponential model is used to model the late reflections component of the echo that corresponds to a multitude of superimposed echo reflections and that typically can be well approximated to be exponentially decaying.

In the embodiments described above, the echo suppression is implemented in a VoIP system (e.g. the received audio signal may include speech of the user 102 for transmission to the user device 110 during a call between the users 102 and 108 over the communication system 100). However, the echo suppression methods described herein can be applied in any suitable system in which echo suppression is to be applied.

In the embodiments described above, and shown in the Figures, echo cancellation (or "echo subtraction") is not applied to the received audio signal y(t). That is, there is no echo cancellation module in the user device 104 and the echo suppression is applied to the received audio signal y(t) without a prior step of applying echo cancellation to the received audio signal y(t).

However, in other embodiments, echo cancellation may be applied, by an echo cancellation module, to the received audio signal y(t). In particular, the echo suppression applied by the echo suppression module 314 may be applied downstream of (i.e. after) the echo cancellation in the processing of the received audio signal y(t). The echo cancellation module would subtract an estimate of the echo signal from the received audio signal, but due to inaccuracies in the estimate of the echo signal, a residual echo would most-likely remain in the received audio signal. It is the residual echo that would then be suppressed by the echo suppression module 314. This echo suppression could be applied in the same way as described herein in the embodiments in which no echo cancellation is applied. If echo subtraction is used, the effect of it can be taken into account in the echo suppression.

The methods described herein may be implemented by executing a computer program product (e.g. the client 206) at the user device 104. That is, a

computer program product may be configured to suppress echo in the received audio signal y(t), wherein the computer program product is embodied on a computer-readable storage medium (e.g. stored in the memory 214) and configured so as when executed on the CPU 202 to perform the operations of any of the methods described herein.

In the methods described above, two models are used: an FIR based model to model the early reflections component of the echo, and an exponential model to model the late reflections component of the echo. However, in other methods, any number (greater than 1) of models may be used to model respective components of the echo. The components of the echo modelled by the different models may, or may not, overlap, such that different models may, or may not, model the same components of the echo as each other. Different ones of the models may be suited to modelling different components of the echo. Each model may be chosen to suit the particular component of the echo for which it is used to model. This can be seen in the example given above in which the FIR based model is suited for modelling the early reflections component of the echo whereas the exponential model is suited for modelling the late reflections component of the echo.

In another example, there may be three models which are respectively suited for, and used for, modelling: (i) the direct sound component of the echo (as shown in Figure 5), (ii) the early reflections component of the echo excluding the direct sound component, and (iii) the late reflections component of the echo. For example, the three models may respectively be: (i) a first FIR based model using a first FIR filter having a fine time spacing between the filter taps such that the model provides a highly accurate model of the direct sound component, (ii) a second FIR based model using a second FIR filter having a time spacing between the filter taps which is coarser than the fine time spacing of the first FIR filter, such that the second FIR based model provides a model of the early reflections component of the echo excluding the direct sound component which is not as accurate as the first FIR based model but for which the complexity is lower for echo paths which are longer than the path of the direct sound component; and (iii) an exponential model which provides the lowest accuracy

of the three models but which also provides the lowest complexity for echo paths having lengths into the late reflections component of the echo.

Generally, any of the functions described herein (e.g. the functional modules shown in Figure 3 and the functional steps shown in Figure 4) can be implemented using software, firmware, hardware (e.g., fixed logic circuitry), or a combination of these implementations. The modules and steps shown separately in Figures 3 and 4 may or may not be implemented as separate modules or steps. For example, the echo suppression module 314 may perform the function of the power estimating modules 306 and 310 and of the combining module 312. The terms "module," "functionality," "component" and "logic" as used herein generally represent software, firmware, hardware, or a combination thereof. In the case of a software implementation, the module, functionality, or logic represents program code that performs specified tasks when executed on a processor (e.g. CPU or CPUs). The program code can be stored in one or more computer readable memory devices. The features of the techniques described herein are platform-independent, meaning that the techniques may be implemented on a variety of commercial computing platforms having a variety of processors.

For example, the user devices may also include an entity (e.g. software) that causes hardware of the user devices to perform operations, e.g., processors functional blocks, and so on. For example, the user devices may include a computer-readable medium that may be configured to maintain instructions that cause the user devices, and more particularly the operating system and associated hardware of the user devices to perform operations. Thus, the instructions function to configure the operating system and associated hardware to perform the operations and in this way result in transformation of the operating system and associated hardware to perform functions. The instructions may be provided by the computer-readable medium to the user devices through a variety of different configurations.

One such configuration of a computer-readable medium is signal bearing medium and thus is configured to transmit the instructions (e.g. as a carrier

wave) to the computing device, such as via a network. The computer-readable medium may also be configured as a computer-readable storage medium and thus is not a signal bearing medium. Examples of a computer-readable storage medium include a random-access memory (RAM), read-only memory (ROM), an optical disc, flash memory, hard disk memory, and other memory devices that may us magnetic, optical, and other techniques to store instructions and other data.

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.

Claims

1. A method of suppressing echo, the method comprising:

outputting an audio signal;

receiving an audio signal, wherein the received audio signal includes echo resulting from said outputted audio signal;

modelling an echo path of the echo in the received audio signal using a plurality of models, wherein a first of the models is a Finite Impulse Response based model, and wherein a second of the models is different to the first model;

using the first model to determine a first model estimate of the echo power of at least a first component of the echo in the received audio signal;

using the second model to determine a second model estimate of the echo power of at least a second component of the echo in the received audio signal;

combining the first and second model estimates of the echo power to determine a combined estimate of the echo power of the echo in the received audio signal; and

using the combined estimate of the echo power to apply echo suppression to the received audio signal, thereby suppressing the echo in the received audio signal.

2. The method of claim 1 wherein the first and second models model different components of the echo in the received audio signal.

3. The method of claim 2 wherein the first model models an early reflections component of the echo in the received audio signal, and the second model models a late reflections component of the echo in the received audio signal.

4. The method of any preceding claim wherein the first model is more accurate than the second model in modelling the echo path of the echo in the received audio signal, and wherein as the length of the echo path increases

the complexity of the first model increases more than the complexity of the second model.

5. The method of any preceding claim wherein said modelling the echo path of the echo in the received audio signal using the first model comprises dynamically adapting a Finite Impulse Response filter estimate $\hat{h}(t)$ in the time domain based on the outputted audio signal and the received audio signal.

6. The method of claim 5 wherein the filter estimate $\hat{h}(t)$ is used to determine the first model estimate of the echo power $P_s^{first}(k,f)$ of said first component of the echo in the received audio signal, according to the equation:

$$\hat{P}_s^{first}(t,f) = \left(\sum_{n=0}^{N} \hat{h}_n(t,f)x(t-n,f)\right)^2,$$

wherein N+1 samples of the outputted audio signal x(t) are considered, and wherein $\hat{h}_n(t)$ are a set of N+1 weights which describe the filter estimate $\hat{h}(t)$.

7. The method of claim 5 wherein the filter estimate $\hat{h}(t)$ is used to determine the first model estimate of the echo power of said first component of the echo in the received audio signal by determining a plurality of power responses from the determined filter estimate $\hat{h}(t)$ by:

partitioning the filter estimate $\hat{h}(t)$ into a plurality of P partitions in the time domain; and

transforming and squaring each of the partitions of the filter estimate $\hat{h}(t)$ to determine a respective power response $|\hat{H}_p(f)|^2$ in the frequency domain for each of the partitions.

8. The method of claim 7 wherein the first model estimate $\hat{P}_s^{first}(k,f)$ of the echo power $P_s^{first}(k,f)$ of said first component of the echo in the received audio signal, for a frame k, is determined according to the equation:

$$\hat{P}_s^{first}(k,f) = \sum_{p=0}^{P-1} |\hat{H}_p(f)|^2 |X(k-p,f)|^2,$$

where $|X(k-p,f)|^2$ is the power spectral density of the outputted signal for frame $k-p$.

9. The method of any preceding claim wherein the second model is an exponential model.

10. The method of claim 9 when dependent upon claim 3 further comprising determining a decay factor estimate $\hat{\gamma}(f)$ of the exponential model based on the outputted audio signal and the received audio signal,

wherein the second model is used to determine the second model estimate $\hat{P}_s^{second}(k,f)$ of the echo power $P_s^{second}(k,f)$ of the second component of the echo in the received audio signal, for a frame k, according to the equation:

$$\hat{P}_s^{second}(k,f) = \hat{\gamma}(f)\hat{P}_{s,end}^{first}(k-1,f) + \hat{\gamma}(f)\hat{P}_s^{second}(k-1,f),$$

where $\hat{P}_{s,end}^{first}(k-1,f)$ is an estimate the echo power of a part of said first component of the echo in the received audio signal determined using said first model for frame $k-1$, said part corresponding to the latest of the reflections in said early reflections component of the echo.

11. The method of claim 10 wherein $\hat{P}_{s,end}^{first}(k-1,f)$ is an estimate of the echo power for the latest of the reflections in said early reflections component determined according to the equation of claim 6 or 8.

12. The method of any preceding claim wherein said combining the first and second model estimates of the echo power comprises summing the first and second model estimates of the echo power.

13. The method of any preceding claim wherein said echo suppression is applied to the received audio signal without a prior step of applying echo cancellation to the received audio signal.

14. The method of any of claims 1 to 12 further comprising applying echo cancellation to the received audio signal, wherein said echo suppression is applied downstream of the echo cancellation in the processing of the received audio signal.

15. The method of any preceding claim wherein the method is performed at a user device for use in a communication event, and wherein the received audio signal comprises speech of a user for transmission from the user device in the communication event.

16. The method of any preceding claim wherein said plurality of models consists of said first and second models only.

17. The method of any of claims 1 to 15 wherein said plurality of models comprises at least one further model in addition to said first and second models.

18. A device configured to implement echo suppression, the device comprising:

audio output apparatus configured to output an audio signal;

audio input apparatus configured to receive an audio signal, wherein the received audio signal includes an echo resulting from said outputted audio signal;

a modelling module configured to model an echo path of the echo in the received audio signal using a plurality of models, wherein the first model is a Finite Impulse Response based model and the second model is different to the first model, wherein the modelling module is configured to use the first model to determine a first model estimate of the echo power of at least a first component of the echo in the received audio signal, and to use the second model to determine a second model estimate of the echo power of at least a second component of the echo in the received audio signal, and wherein the modelling module comprises a combining module configured to combine the first and second model estimates of the echo power to determine a combined estimate of the echo power of the echo in the received audio signal; and

an echo suppression module configured to use the combined estimate of the echo power to apply echo suppression to the received audio signal, thereby suppressing the echo in the received audio signal.

19. The device of claim 18 wherein the audio output apparatus comprises a speaker configured to output the outputted audio signal, and wherein the audio input apparatus comprises a microphone configured to receive the received audio signal.

20. A computer program product configured to suppress echo in a received audio signal, the computer program product being embodied on a computer-readable storage medium and configured so as when executed on a processor to perform the operations of any of claims 1 to 17.

Intellectual Property Office

Application No:	GB1223238.5	**Examiner:**	Dr Mark Lewney
Claims searched:	1-20	**Date of search:**	29 May 2014

Patents Act 1977: Search Report under Section 17

Documents considered to be relevant:

Category	Relevant to claims	Identity of document and passage or figure of particular relevance
X	1-5 & 12-20	WO2006/040734 A1 (PHILIPS) - See especially figs. 4, 9 & 10 and paragraphs [0011 & 0054-65].
X	1-5 & 12-20	US7068780 B1 (LEVONAS ET AL.) - See especially figs. 3 & 5 and accompanying description.
X	1-5 & 12-20	US6256383 B1 (CHEN) - See especially fig. 1 and accompanying description.

Categories:

X	Document indicating lack of novelty or inventive step	A	Document indicating technological background and/or state of the art.
Y	Document indicating lack of inventive step if combined with one or more other documents of same category.	P	Document published on or after the declared priority date but before the filing date of this invention.
&	Member of the same patent family	E	Patent document published on or after, but with priority date earlier than, the filing date of this application.

Field of Search:

Search of GB, EP, WO & US patent documents classified in the following areas of the UKC^{X} :

Worldwide search of patent documents classified in the following areas of the IPC

G10L; H04B; H04M

The following online and other databases have been used in the preparation of this search report

WPI, EPODOC.

International Classification:

Subclass	Subgroup	Valid From
G10L	0021/0208	01/01/2013
H04B	0003/20	01/01/2006
H04M	0009/08	01/01/2006