(12) PATENT
(19) AUSTRALIAN PATENT OFFICE

(11) Application No. AU 199646721 B2
(10) Patent No. 707896

(54) Title
Voice activity detection

(51)6 International Patent Classification(s)
G10L 003/00

(21) Application No: 199646721 (22) Application Date: 1996 .02 .15

(87) WIPO No: WO96/25733

(30) Priority Data

(31) Number	(32) Date	(33) Country
95300975	1995 .02 .15	EP

(43) Publication Date : 1996 .09 .04
(43) Publication Journal Date : 1996 .10 .24
(44) Accepted Journal Date : 1999 .07 .22

(71) Applicant(s)
British Telecommunications public limited company

(72) Inventor(s)
James Anthony Bridges

(74) Agent/Attorney
BALDWIN SHELSTON WATERS,Level 21,60 Margaret Street,SYDNEY NSW 2000

(56) Related Art
US 5155760
US 4410763

OPI DATE 04/09/96 APPLN. ID 46721/96
AOJP DATE 24/10/96 PCT NUMBER PCT/GB96/00344

AU9646721

CT)

(51) International Patent Classification 6 :
G10L 3/00

A1

(11) International Publication Number: **WO 96/25733**

(43) International Publication Date: 22 August 1996 (22.08.96)

(21) International Application Number: PCT/GB96/00344

(22) International Filing Date: 15 February 1996 (15.02.96)

(30) Priority Data:
95300975.0 15 February 1995 (15.02.95) EP
(34) Countries for which the regional or international application was filed: GB et al.

(71) Applicant *(for all designated States except US)*: BRITISH TELECOMMUNICATIONS PUBLIC LIMITED COMPANY [GB/GB]; 81 Newgate Street, London EC1A 7AJ (GB).

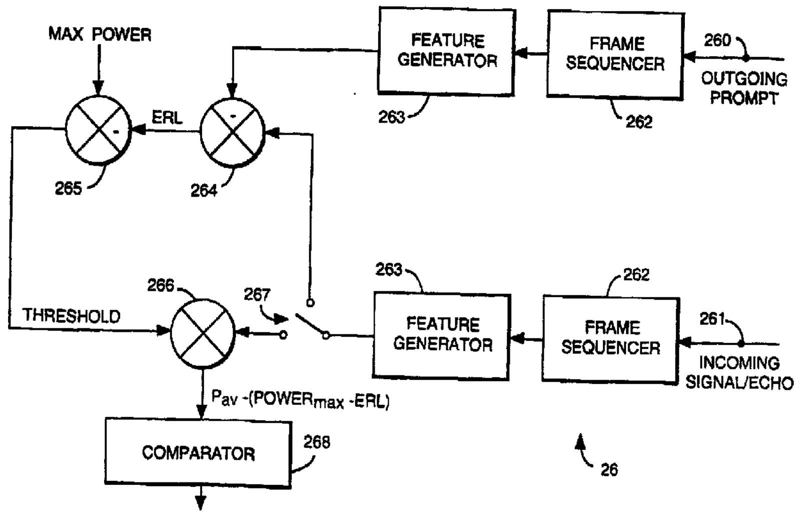
(72) Inventor; and
(75) Inventor/Applicant *(for US only)*: BRIDGES, James, Anthony [GB/GB]; 88 Suffolk Road, Ipswich, Suffolk IP4 2EL (GB).

(74) Agent: HIBBERT, Juliet, Jane, Grace; BT Group Legal Services, Intellectual Property Dept., 8th floor - Holborn Centre, 120 Holborn, London EC1N 2TE (GB).

(81) Designated States: AL, AM, AT, AU, AZ, BB, BG, BR, BY, CA, CH, CN, CZ, DE, DK, EE, ES, FI, GB, GE, HU, IS, JP, KE, KG, KP, KR, KZ, LK, LR, LS, LT, LU, LV, MD, MG, MK, MN, MW, MX, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, TJ, TM, TR, TT, UA, UG, US, UZ, VN, ARIPO patent (KE, LS, MW, SD, SZ, UG), Eurasian patent (AZ, BY, KG, KZ, RU, TJ, TM), European patent (AT, BE, CH, DE, DK, ES, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, ML, MR, NE, SN, TD, TG).

Published
With international search report.

(54) Title: VOICE ACTIVITY DETECTION



(57) Abstract

A voice activity detector (26) comprising an input for receiving an outgoing speech signal transmitted from a speech system (2) to a user and an input for receiving an incoming signal from the user. Both the outgoing and incoming signals are divided into time limited frames. Means (263) are provided for calculating a feature from each frame of the incoming signal and for forming a function of the calculated feature and a threshold. Based on the function, it is determined whether or not the incoming signal includes speech. Means are provided to determine the echo return loss during an outgoing speech signal from the interactive speech system and to control the threshold in dependence on the echo return loss measured.

VOICE ACTIVITY DETECTION

This invention relates to voice activity detection.

There are many automated systems that depend on the detection of speech for operation, for instance automated speech systems and cellular radio coding systems. Such systems monitor transmission paths from users' equipment for the occurrence of speech and, on the occurrence of speech, take appropriate action. Unfortunately transmission paths are rarely free from noise. Systems which are arranged simply to detect activity on the path may therefore incorrectly take action if there is noise present.

The usual noise that is present is line noise (i.e. noise that is present irrespective of whether or not the a signal is being transmitted) and background noise from a telephone conversation, such as a dog barking, the sound of the television, the noise of a car's engine etc.

Another source of noise in communications systems is echo. For instance, echoes in a public switch telephone network (PSTN) are essentially caused by electrical and/or acoustic coupling e.g. at the four wire to two wire interface of a conventional exchange box; or the acoustic coupling in a telephone handset, from earpiece to microphone. The acoustic echo is time variant during a call due to the variation of the airpath, i.e. the talker altering the position of their head between the microphone and the loudspeaker. Similarly in telephone kiosks, the interior of the kiosk has a limited damping characteristic and is reverberant which results in resonant behaviour. Again this causes the acoustic echo path to vary if the talker moves around the kiosk or indeed with any air movement. Acoustic echo is becoming a more important issue at this time due to the increased use of hands free telephones. The effect of the overall echo or reflection path is to attenuate, delay and filter a signal.

The echo path is dependent on the line, switching route and phone type. This means that the transfer function of the reflection path can vary between calls since any of the line, switching route and the handset may change from call to call as different switch gear will be selected to make the connection.

Various techniques are known to improve the echo control in human-to-human speech communications systems. There are three main techniques. Firstly insertion losses may be added into the talker's transmission path to reduce the level of the outgoing signal. However the insertion losses may cause the received signal to become intolerably low for the listener. Alternatively, echo suppressors operate on the principle of detecting signal levels in the transmitting and receiving path and then comparing the levels to determine how to operate switchable insertion loss pads. A high attenuation is placed in the transmit path when speech is detected on the received path. Echo suppressors are usually used on longer delay connections such as international telephony links where suitable fixed insertion losses would be insufficient.

Echo cancellers are voice operated devices which use adaptive signal processing to reduce or eliminate echoes by estimating an echo path transfer function. An outgoing signal is fed into the device and the resulting output signal subtracted from the received signal. Provided that the model is representative of the real echo path, the echo should theoretically be cancelled. However, echo cancellers suffer from stability problems and are computationally expensive. Echo cancellers are also very sensitive to noise bursts during training.

One example of an automated speech system is the telephone answering machine, which records messages left by a caller. Generally, when a user calls up an automated speech system, a prompt is played to the user which prompt usually requires a reply. Thus an outgoing signal from the speech system is passed along a transmission line to the loudspeaker of a user's telephone. The user then provides a response to the prompt which is passed to the speech system which then takes appropriate action.

It has been proposed that allowing a caller to an automated speech system to interrupt outgoing prompts from the system greatly enhances the usability of the system for those callers who are familiar with the dialogue of the system. This facility is often termed "barge in" or "over-ridable guidance".

If a user speaks during a prompt, the spoken words may be preceded or corrupted by an echo of the outgoing prompt. Essentially isolated clean vocabulary utterances from the user are transformed into embedded vocabulary utterances (in which the vocabulary word is contaminated with additional sounds). In automated speech systems which involve automated speech recognition, because of the limitations of current speech recognition technology, this results in a reduction in recognition performance.

If a user has never used the service provided by the automated speech system, the user will need to hear the prompts provided by the speech generator in their entirety. However, once a user has become familiar with the service the information that is required at each stage, the user may wish to provide the required response before the prompt is finished. If a speech recogniser or recording means is turned off until the prompt is finished, no attempt will be made to recognise a user's early response. If, on the other hand, the speech recogniser or recording means is turned on all the time, the input would include both the echo of the outgoing prompt and the response provided by the user. Such a signal would be unlikely to be recognisable by a speech recogniser. Voice activity detectors (VADs) have therefore been developed to detect voice activity on the path.

Known voice activity detectors rely on generating an estimate of the noise in an incoming signal and comparing an incoming signal with the estimate which is either fixed or updated during periods of non-speech. An example of such a voice activated system is described in US Patent No. 5155760 and US Patent No. 4410763.

Voice activity detectors are used to detect speech in the incoming signal, and to interrupt the outgoing prompt and turn on the recogniser when such speech is detected. A user will hear a clipped prompt. This is satisfactory if the user has barged in. If however the voice activity detector has incorrectly detected speech, the user will hear a clipped prompt and have no instructions on to how to proceed with the system. This is clearly undesirable.

The present invention provides an interactive speech apparatus comprising:

a speech generator for generating an outgoing speech signal; and

a voice activity detector comprising:

an input for receiving said outgoing speech signal;

an input for receiving incoming echo and speech signals;

means arranged in operation to derive, during the beginning of said outgoing speech signal, the echo return loss from the difference in the level of said outgoing speech signal and the level of the echo thereof;

means arranged in operation to calculate a threshold in dependence on said echo return loss;

means arranged in operation to evaluate a function of one of a plurality of features calculated from respective frames of said incoming signal and said threshold;

means arranged to determine, based on the evaluation, whether or not the incoming signal includes direct speech from a user of the apparatus; and

means arranged to control the operation of said speech apparatus responsive to the detection of direct speech from the user.

The echo return loss is a measure of attenuation of the outgoing prompt by the transmission path.

Controlling the threshold on the basis of the echo return loss measured not only reduces the number of false triggering by the voice activity detector due to echo, but also reduces the number of triggerings of the voice activity detector when the user makes a response over a line having a high amount of echo. Whilst this may appear unattractive, it should be appreciated that it is preferable for the voice activity detector not to trigger when the user barges in than for the voice activity detector to trigger when the user has not barged in, which would leave the user with a clipped prompt and no further assistance.

The threshold may be a function of the echo return loss and the maximum possible power of the outgoing signal. Both of these are long-term characteristics of the line (although the echo return loss may be remeasured from time to time). Preferably the threshold is the difference between the maximum power and the echo return loss. It may be preferred that the threshold is a function of the echo return loss and the feature calculated from each frame of the outgoing speech signal (i.e. the threshold represents an attenuation of each frame of the outgoing signal).

Preferably the feature calculated is the average power of each frame of a signal although other features, such as the frame energy, may be used. More than one feature of the incoming signal may be calculated and various functions formed.

The voice activity detector may further include data relating to statistical models representing the calculated feature for at least a signal containing substantially noise-free speech and a noisy signal, the function of the calculated feature and the threshold being compared with the statistical models. The noisy signal statistical models may represent line noise and/or typical background noise and/or an echo of the outgoing signal.

In accordance with the invention there is provided a method of operating an interactive speech apparatus, said method comprising the steps of:

transmitting an outgoing speech prompt signal to a user;

receiving an incoming echo signal;

deriving, during the beginning of said outgoing speech signal, the echo return loss from the difference in the level of the outgoing speech signal and the level of the echo thereof;

calculating a threshold in dependence on said echo return loss;

evaluating a function of a feature of the incoming signal and said threshold;

detecting a user's spoken response to said prompt on the basis of said evaluation; and

controlling the operation of said interactive speech apparatus responsive to the detection of the user's spoken response.

Preferably the threshold is a function of the echo return loss and the maximum possible power of the outgoing signal. As mentioned above, the threshold may be a function of the echo return loss and the same feature calculated from a frame of the outgoing speech signal. The feature calculated may be the average power of each frame of a signal.

Unless the context clearly requires otherwise, throughout the description and the claims, the words 'comprise', 'comprising', and the like are to be construed in an inclusive sense as opposed to an exclusive or exhaustive sense; that is to say, in the sense of "including, but not limited to".

The invention will now be further described by way of example with reference to the accompanying drawings in which:

Figure 1 shows an automated speech system including a voice activity detector according to the invention; and

Figure 2 shows the components of a voice activity detector according to the invention.

Figure 1 shows an automated speech system 2, including a voice activity detector according to the invention, connected via the public switched telephone network to a user terminal, which is usually a telephone 4. The automated speech system is preferably located at an exchange in the network. The automated speech system 2 is connected to a hybrid transformer 6 via an outgoing line 8 and an incoming line 10. A user's telephone is connected to the hybrid via a two-way line 12.

Echoes in the PSTN are essentially caused by electrical and/or acoustic coupling e.g., the four wire to two wire interface at the hybrid transformer 6 (indicated by the arrow 7). Acoustic coupling in the handset of the telephone 4, from earpiece to microphone, causes acoustic echo (indicated by the arrow 9).

The automated speech system 2 comprises a speech generator 22, a speech recogniser 24 and a voice activity detector (VAD) 26. The type of speech generator 22 and speech recogniser 24 will not be discussed further since these do not form part of the invention. It will be clear to a person skilled in the art that any suitable speech generator, for instance those using text to speech technology or pre-recorded messages, may be used. In addition any suitable type of speech recogniser 24 may be used.

In use, when a user calls up the automated speech system the speech generator 22 plays a prompt to the user, which usually requires a reply. Thus an outgoing speech signal from the speech system is passed along the transmission line 8 to the hybrid transformer 6 which switches the signal to the loudspeaker of the user's telephone 4. At the end of a prompt, the user provides a response which is passed to the speech recogniser 24 via the hybrid 6 and the incoming line 10. The speech recogniser 24 then attempts to recognise the response and appropriate action is taken in response to the recognition result.

If a user has never used the service provided by the automated speech system, the user will need to hear the prompts provided by the speech generator 22 in their entirety. However, once a user has become familiar with the service and the information that is required at each stage, the user may wish to provide the required response before the prompt has finished. If the speech recogniser 24 is turned off until the prompt is finished, no attempt will be made to recognise the user's early response. If, on the other hand, the speech recogniser 24 is turned on all the time, the input to the speech recogniser would include both the echo of the outgoing prompt and the response provided by the user. Such a signal would be unlikely to be recognisable by the speech recogniser.

The voice activity detector 26 is provided to detect direct speech (i.e. speech from the user) in the incoming signal. The speech recogniser 24 is held in an inoperative mode until speech is detected by the voice activity detector 26. An output signal from the voice activity detector 26 passes to the speech generator 22, which is then interrupted (so clipping the prompt), and the speech recogniser 24, which, in response, becomes active.

Figure 2 shows the voice activity detector 26 of the invention in more detail. The voice activity detector 26 has an input 260 for receiving an outgoing prompt signal from the speech generator 22 and an input 261 for receiving the signal received via the incoming line 10. For each signal, the voice activity detector includes a frame sequencer 262 which divides the incoming signal into frames of data comprising 256 contiguous samples. Since the energy of speech is relatively stationary over 15 milliseconds, frames of 32 ms are preferred with an overlap of 16 ms between adjacent frames. This has the effect of making the VAD more robust to impulsive noise.

The frame of data is then passed to a feature generator 263 which calculates the average power of each frame. The average power of a frame of a signal is determined by the following equation:

$$\text{Log Average Frame Power } P_{av} = 10 \log_{10} \frac{\sum_{n=1}^{N} f_n(t)^2}{N}$$

where N is the number of samples in a frame, in this case 256.

Echo return loss is a measure of the attenuation i.e. the difference (in decibels) between the outgoing and the reflected signal. The echo return loss (ERL) is the difference between features calculated for the outgoing prompt and the returning echo i.e.

$$\text{ERL} = 10 \log_{10} \left[\frac{1}{N} \sum_{i=1}^{N} P_i(t) \middle| \text{outgoing prompt} \right] - 10 \log_{10} \left[\frac{1}{N} \sum_{i=1}^{N} P_i(t) \middle| \text{incoming echo} \right]$$

where N is the number of samples over which the average power P_i is calculated. N should be as high as is practicable.

As can be seen from Figure 2, the echo return loss is determined by subtracting the average power of a frame of the incoming echo from the average power of a frame of the outgoing prompt. This is achieved by exciting the transmission path 8, 10 with a prompt from the system, such as a welcome prompt. The signal level of the outgoing prompt and the returning echo are then calculated as described above by frame sequence 262 and feature generator 263. The resulting signal levels are subtracted by subtractor 264 to form the echo return loss.

The echo return loss is then subtracted by subtractor 265 from the maximum power possible for the transmission path i.e. the subtractor 265 calculates the threshold signal:

$$Threshold = Maximum\ possible\ power - echo\ return\ loss$$

Typical echo return loss is approximately 12dB although the range is of the order of 6-30dB the maximum possible power on a telephone line for an A-law signal is around 72dB.

The ERL is calculated from the first 50 or so frames of the outgoing prompt, although more or fewer frames may be used.

Once the ERL has been calculated, the switch 267 is switched to pass the data relating to the incoming lime to the subtractor 266. The threshold signal is then, during the remainder of the call, subtracted by subtractor 266 from the average power of each frame of the incoming signal. Thus the output of the subtractor 266 is

$$P_{av}|_{\text{incoming signal}} - (\text{Max } possible\ power - ERL)$$

The output of subtractor 266 is passed to a comparator 268, which compares the result with a threshold. If the result is above the threshold, the incoming signal is deemed to include direct speech from the user and a signal is output from the voice activity detector to deactivate the speech generator 22 and activate the speech recogniser 24. If the result is lower than the threshold, no signal is output from the voice activity detector and the speech recogniser remains inoperative.

In another embodiment of the invention, the output of subtractor 266 is passed to a classifier (not shown) which classifies the incoming signal as speech or non-speech. This may be achieved by comparing the output of subtractor 266 with statistical models representing the same feature for typical speech and non-speech signals.

In a further embodiment, the threshold signal is formed according to the following equation:

$$(P_{av}|_{\text{outgoing prompt}} - ERL)$$

The resulting threshold signal is input to subtractor 266 to form the product:

$$P_{av}|_{\text{incoming signal}} - (P_{av}|_{\text{outgoing prompt}} - ERL)$$

The echo return loss is calculated at the beginning of at least the first prompt from the speech system. The echo return loss can be calculated from a single frame if necessary, since the echo return loss is calculated on a frame-by-frame basis. Thus, even if a user speaks almost immediately it is still possible for the echo return loss to be calculated.

The frame sequencers 262 and feature generators 263 have been described as being an integral part of the voice activity detector. It will be clear to a skilled person that this is not an essential feature of the invention, either or both of these being separate components. Equally it is not necessary for a separate frame sequencer and feature generator to be provided for each signal. A single frame sequencer and feature generator may be sufficient to generate a feature from each signal.

THE CLAIMS DEFINING THE INVENTION ARE AS FOLLOWS:-

1. An interactive speech apparatus comprising:

a speech generator for generating an outgoing speech signal; and

a voice activity detector comprising:

an input for receiving said outgoing speech signal;

an input for receiving incoming echo and speech signals;

means arranged in operation to derive, during the beginning of said outgoing speech signal, the echo return loss from the difference in the level of said outgoing speech signal and the level of the echo thereof;

means arranged in operation to calculate a threshold in dependence on said echo return loss;

means arranged in operation to evaluate a function of one of a plurality of features calculated from respective frames of said incoming signal and said threshold;

means arranged to determine, based on the evaluation, whether or not the incoming signal includes direct speech from a user of the apparatus; and

means arranged to control the operation of said speech apparatus responsive to the detection of direct speech from the user.

2. An interactive speech apparatus according to claim 1 wherein the threshold is a function of the echo return loss and the maximum possible power of the outgoing signal.

3. An interactive speech apparatus according to claim 1 wherein the threshold is a function of the echo return loss and a feature calculated from a frame of the outgoing speech signal.

4. An interactive speech apparatus according to any of claims 1, 2 or 3 wherein the feature calculated is the average power of each frame of a signal.

5. A method of operating an interactive speech apparatus, said method comprising the steps of:

transmitting an outgoing speech prompt signal to a user;

receiving an incoming echo signal;

deriving, during the beginning of said outgoing speech signal, the echo return loss from the difference in the level of the outgoing speech signal and the level of the echo thereof;

calculating a threshold in dependence on said echo return loss;

evaluating a function of a feature of the incoming signal and said threshold;

detecting a user's spoken response to said prompt on the basis of said evaluation; and

controlling the operation of said interactive speech apparatus responsive to the detection of the user's spoken response.

6. A method according to claim 5 wherein the threshold is a function of the echo return loss and the maximum possible power of the outgoing signal.

7. A method according to claim 5 wherein the threshold is a function of the echo return loss and the same feature calculated from a frame of the outgoing speech signal.

8. A method according to any of claims 5 to 7 wherein the feature calculated is the average power of each frame of a signal.

9. An interactive speech apparatus substantially as herein described with reference to Figure 2 of the accompanying drawings.

10. A method of operating an interactive speech apparatus substantially as herein described with reference to Figure 2 of the accompanying drawings.

DATED this 3rd day of MAY, 1999

BRIITISH TELECOMMUNICATIONS public limited company

Attorney: PETER R. HEATHCOTE
Fellow Institute of Patent Attorneys of Australia
of BALDWIN SHELSTON WATERS

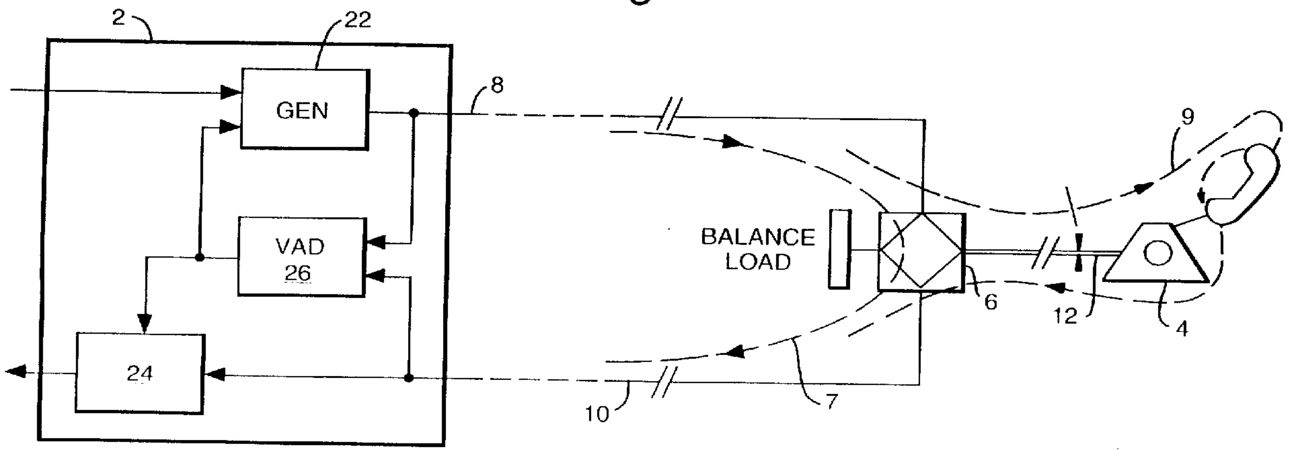
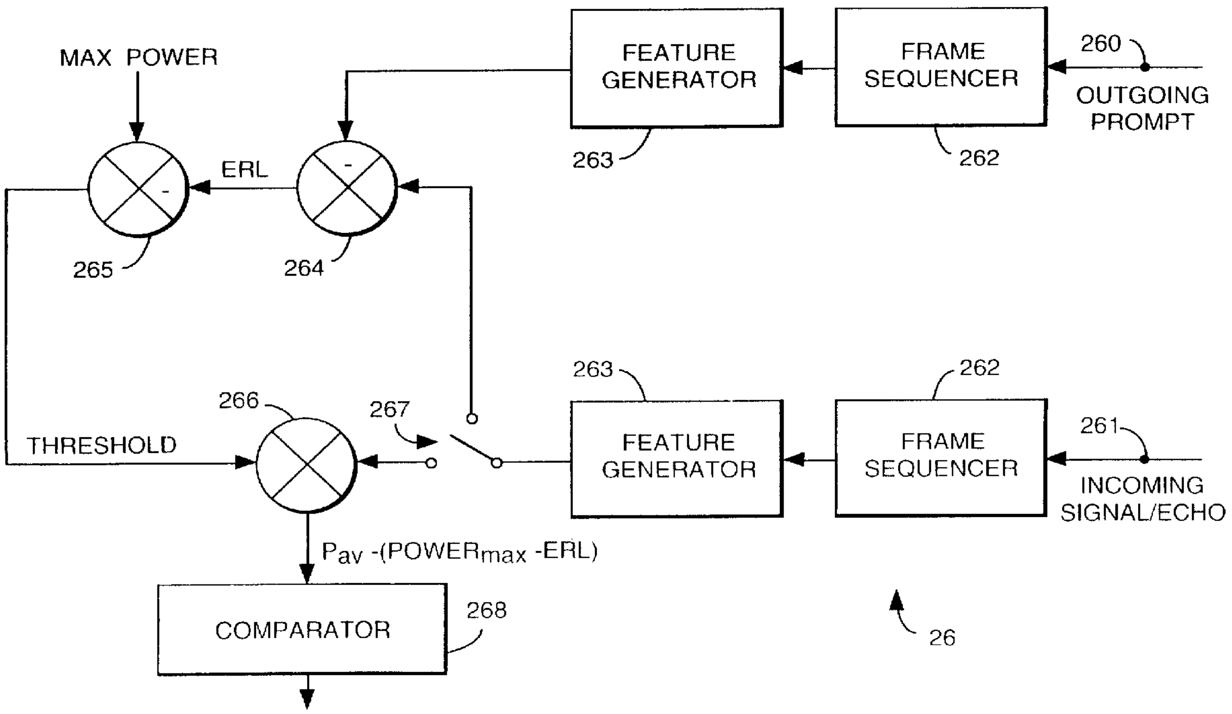


Fig.1.

Fig.2.


260
OUTGOING PROMPT
FRAME SEQUENCER
262
FEATURE GENERATOR
263
261
INCOMING SIGNAL/ECHO
FRAME SEQUENCER
262
FEATURE GENERATOR
263
26
MAX POWER
ERL
-
-
264
265
267
266
THRESHOLD
P_{av} -(POWER$_{max}$ -ERL)
COMPARATOR
268