

A NOISE-STRIPPING DEVICE

166166

Field Of Invention

The invention relates generally to speech processing. In particular, the invention relates to a noise-stripping device for speech processing.

Background

The use of noise-stripping techniques for improving speech intelligibility is widely known and practiced in the field of speech processing. Typically, conventional noise-stripping techniques involve gain modification of different spectral regions of speech signals representative of articulated speech, and the degree of gain modification applied to any spectral region of speech signals depends on the signal-to-noise ratio (SNR) of that spectral region. A number of conventional noise-stripping techniques are disclosed in patents. Each of these techniques when applied to speech processing to a limited degree reduces noise in noise-contaminated speech signals, but does so usually at the expense of speech quality. The effectiveness of such techniques also lessens with increasing noise levels in the noise-contaminated speech signals.

A common problem that exists amongst the conventional noise-stripping techniques is the proper identification of speech and background noise in speech captured or recorded in a noisy environment. In such situations, speech is captured or recorded together and mixed with the background noise, therefore resulting in noise-contaminated speech signals. Since speech and background noise have not been properly identified in such noise-contaminated speech signals, the task of performing gain modification thereon for isolating uncontaminated speech signals is usually minimally successful.

A number of US patents teach or disclose noise-stripping techniques, but such teachings or disclosures have not been applied with satisfactory results. These patents include US Patent 4811404 by Vilmur et al, US Patent 6001131 by Raman, and US Patents 4628529 and 4630305 by Borth et al.



G00002

Vilmur et al, incorporating Borth et al (4628529), discloses a noise-stripping technique that applies spectral subtraction, or spectral gain modification, for enhancing speech quality in which gain modification is performed on noise-contaminated speech signals by limiting gain in particular spectral regions or channels of a noise-contaminated speech signal that do not reach a specified SNR threshold. A voice metric calculator provides measurements of voice-like characteristics of a channel by measuring the SNR of the channel and using the SNR for obtaining a corresponding voice metric value from a preset table. The voice metric value is then used to determine if background noise is present in the channel by comparing such a value with a predetermined threshold value. The voice metric calculator also determines the length of time intervals between updates of background noise values relating to the channel, such information being used to determine gain factors for gain modification to the channel.

Raman discloses a technique that relies on identifying ambient noise in noise-contaminated speech signals following a predetermined duration of speech signals as a basis for noise cancellation by using a speech/noise distinguishing threshold.

Borth et al (4630305) teaches a technique which involves splitting noise-contaminated speech signals into channels and using an automatic channel gain selector for controlling channel gain depending on the SNR of each channel. Channel gain is selected automatically from a preset gain table by reference to channel number, channel SNR, and overall background noise level of the channel.

There is therefore clearly a need for a background noise-stripping device and a corresponding method for identifying speech and background noise in noise-contaminated speech, thereafter processing the same for retrieving the speech.

Summary

In accordance with a first aspect of the invention, a method for stripping background noise component from a noise-contaminated speech signal is provided, the method comprising the steps of:

digitising the noise-contaminated speech signal to form samples grouped into frames;

dividing in the frequency domain the digitised signal into a plurality of frequency bins;

storing a plurality of frames of digitised signal equivalent to a preset length of digitised signal in a buffer;

estimating the spectrum level of a current frame of digitised signal during a preset period;

comparing the spectrum estimate of the current frame of digitised signal with a spectrum estimate representative of an earlier frame of digitised signal and selecting the lower of the two spectrum estimates during the preset period;

storing the selected lower spectrum estimate in the buffer during the preset period;

assigning the stored and selected lower spectrum estimate as representative of the current frame of digitised signal; and

setting as background noise spectrum estimate the minimum value of the stored and selected lower spectrum estimates of the plurality of frames stored in the buffer.

In accordance with a second aspect of the invention, a device for stripping background noise component from a noise-contaminated speech signal is provided, the device comprising:

means for digitising the noise-contaminated speech signal to form samples grouped into frames;

means for dividing in the frequency domain the digitised signal into a plurality of frequency bins;

means for storing a plurality of frames of digitised signal equivalent to a preset length of digitised signal in a buffer;

means for estimating the spectrum level of a current frame of digitised signal during a preset period;

means for comparing the spectrum estimate of the current frame of digitised signal with a spectrum estimate representative of an earlier frame of digitised signal and selecting the lower of the two spectrum estimates during the preset period;

means for storing the selected lower spectrum estimate in the buffer during the preset period;

means for assigning the stored and selected lower spectrum estimate as representative of the current frame of digitised signal; and

means for setting as background noise spectrum estimate the minimum value of the stored and selected lower spectrum estimates of the plurality of frames stored in the buffer.

Brief Description Of The Drawings

Embodiments of the invention are described in detail hereafter with reference to the drawings, in which:

Figure 1 provides a block diagram showing modules in a noise-stripping device according to a first embodiment of the invention implemented using a fixed-point processor;

Figure 2 provides a block diagram showing modules in a noise-stripping device according to a second embodiment of the invention implemented using a floating-point processor;

Figure 3 provides a block diagram showing calculation steps for estimation of spectrum relating to background noise;

Figure 4 provides a block diagram showing steps performed in a gain modification process in respective modules in a gain vector modification module in the floating-point device of Figure 2; and

Figure 5 provides a block diagram showing a gain modification process for the fixed-point device of Figure 1.

Detailed Description

In applying improved noise-stripping techniques involving spectral subtraction described hereinafter, noise-stripping devices according to embodiments of the invention afford the advantage of enhancing speech intelligibility in the presence of background noise. An

application of such a device is in the field of enhancing speech clarity for performing automatic voice switching.

Conventional noise-stripping techniques are limited in the ability to properly identify speech and background noise components of signals representing speech contaminated with background-noise when substantially removing or reducing the background noise components from the noise-contaminated speech signals. Also, particular noise-stripping processes used in these techniques introduce artifacts and distort speech.

While conventional techniques rely on thresholds to make speech/noise decisions and/or identification of speech components for quantifying noise components following the speech components, the noise-stripping devices according to embodiments of the invention place emphasis on the identification of noise components. Most human speech patterns show that every 0.5 to 1 second of articulated speech is typically interspersed with at least one non-voice pause, during which background noise may be isolated, while most noise patterns do not show such periodic behaviour. The devices identify background noise during pauses in speech and accordingly adjust gain vectors for eliminating the background noise with minimum distortion of speech.

Algorithms are also applied in the noise-stripping devices for the characterization of background noise and for gain adjustment of background noise and speech components of a captured or recorded noise-contaminated speech signal.

In the noise-stripping devices of which processing modules are shown in Figures 1 and 2, a noise-contaminated speech signal is preferably sampled and digitised at 16kHz into samples of the noise-contaminated speech signal with 128 samples constituting a frame, so that digital signal processing may be applied. Any type of digital signal processors, combination of digital signal processing elements, or computer-aided processors or processing elements capable of processing digital signals, performing digital signal processing, or in general carrying out computations or calculations in accordance with formulas or equations, may be used in the device. Processing steps, calculations, procedures, and generally processes may be performed in modules or the like components that may be independent processing elements or parts of a processor, so that

these processing elements may be implemented by way of hardware, software, firmware, or combination thereof.

The frame in the time domain is applied time-based processing and analysis by the noise-stripping devices, and converted to the frequency domain preferably using Fast Fourier Transform (FFT) techniques for frequency-based processing and analysis. Each frame in the frequency domain is divided into narrow frequency bands known as FFT bins, whereby each FFT bin is preferably set to 62.5Hz in width. For eventual gain modification, the digitised signals are preferably processed independently in different spectral regions, of which values are preferably specified that include the bass (<1250Hz), mid-frequency (1250-4000Hz) and high-frequency (>4000Hz) spectral regions.

The operational aspects of the noise-stripping devices are described hereinafter in greater detail with reference to Figures 1 and 2. During operation the noise-stripping devices digitise a noise-contaminated signal from a microphone or the like pick-up transducer and provide the digitised signal to a digital signal processor in which the background noise component is substantially removed or reduced. The speech-enhanced signal is then converted to an analog output.

Fixed- and floating-point processors are used respectively in noise-stripping devices shown in Figures 1 and 2, with a number of processing modules differing as shown therein. Fixed-point processors have lower power consumption and are favoured for many portable applications. However, a number of processing steps described hereinafter in relation to the floating-point implementation are not included in the fixed-point implementation due to a possibility of overflow that affects the dynamic range in the fixed-point processor in respect of FFT processing. Floating-point processors are therefore more powerful and provide better noise reduction and speech quality in respect of the current intents and purposes. For example, the process of windowing alone used in the fixed-point implementation reduces aliasing distortion, albeit not as effectively as the combined processes of windowing and gain vector rotation and truncation used in the floating-point implementation.

As shown in Figures 1 and 2, a noise-contaminated speech signal is first input to and processed by an Analog-to-Digital (A/D) Converter 12 for conversion into a digital signal consisting of frames of samples. In the fixed-point implementation in Figure 1, the A/D Converter 12 outputs the digital signal to an Emphasis Filter 14 (of first order FIR filter) for enhancing high frequency elements of the speech component.

The Emphasis Filter 14 in the fixed-point device or the A/D Converter 12 in the floating-point device provides input to a Frame Overlap & Window module 16 in which the input consisting of two frames, i.e. a current frame and a previous frame, is overlapped and processed using a windowing function to form a windowed current block of samples consisting of 256 samples for subsequent FFT operation. The processing of such a block, until the retrieval of the current frame performed in an Overlap yfrain module 40 described hereinafter, involves both the current and previous frames although the current frame remains the frame of interest during the description hereinafter. To retrieve the current frame, samples in the previous frame from the windowed current block and samples in a current frame from a windowed previous block are added to form the output of the Overlap yfrain module 40. This is possible because by applying a symmetric windowing technique in the Frame Overlap & Window module 16, in which windowed blocks are symmetrical about central points, the addition of the current and previous blocks in Overlap yfrain module 40 yields the current frame. The symmetric windowing technique is, for example, a Hamming windowing technique or the preferred Hanning windowing technique.

However for purposes of simplicity and brevity, when any reference is hereinafter made to the current frame of sample until the retrieval of the current frame in the Overlap yfrain module 40, such reference is made to the current block of samples, which for all intents and purposes, includes the current frame of samples.

The output of the Frame Overlap & Window module 16 is provided as input to an FFT module 18 for conversion to the frequency domain for further processing. The current frame of samples after conversion to the frequency is defined as an output Xffts, in which the first 129 bins are used as a calculation frame in frequency domain.

The magnitude or power spectrum S relating to the current calculation frame of the input noise-contaminated speech signal, which consists of both speech and background noise components, is calculated using the first 129 bins of the frequency domain output X_{fft} in a spectrum calculation module 20. In this module, the magnitude calculation operation is performed on the first 129 bins of X_{fft} to provide the magnitude spectrum of the current calculation frame in the fixed-point implementation in Figure 1, and a magnitude squaring operation performed on the first 129 bins of X_{fft} to provide the power spectrum of the current calculation frame in the floating-point implementation in Figure 2.

Next an estimation of the spectrum relating to the input noise-contaminated speech signal is performed in a signal-plus-noise spectrum estimation module 22. The signal-plus-noise spectrum estimation module 22 first averages the magnitude or power spectrum S over three to five calculation frames of the input noise-contaminated speech signal, then calculates the estimation of the spectrum S_c relating to the input noise-contaminated speech signal using equation (1)

Firstly:

$$D(i) = \sum_{j=1}^k S(i, j); \quad k=3 \sim 5, i=0, \dots, N;$$

where S is the power spectrum relating to a calculation frame of input noise-contaminated speech signal consisting of both speech and background noise components processed in the floating-point implementation in Figure 2, or the magnitude spectrum relating to the calculation frame of noise-contaminated input signal processed in the fixed-point implementation Figure 1; i is the FFT bin number; N is the maximum order of a calculation frame; and $D(i)$ is the value of $S(i)$ averaged over k frames.

Then:

$$S_c(b) = \frac{1}{nav} \sum_{i=1}^{nav} D(i); \quad \text{for } i = b, \dots, b + nav, 0 \leq b \leq N, \quad (1)$$

in which $\text{nav} = \begin{cases} 0 & \text{for } f(b) < 1000\text{Hz} \\ BW / B1 & \text{for } f(b) \geq 1000\text{Hz} \end{cases}$ and $D(N) = D(i)$, for $i > N$,

where:

S_c is an estimation of the spectrum relating to the input noise-contaminated speech signal;

b, i is the FFT bin number;

$f(b)$ is the frequency of FFT bin b ;

$B1$ is the width of the FFT bin;

and preferably

$BW = 150\text{ Hz}$ for $1000\text{Hz} \leq f(b) < 1500\text{Hz}$;

$BW = 250\text{ Hz}$ for $1500\text{Hz} \leq f(b) < 2000\text{Hz}$;

$BW = 350\text{ Hz}$ for $2000\text{Hz} \leq f(b) < 3000\text{Hz}$;

$BW = 500\text{ Hz}$ for $3000\text{Hz} \leq f(b) < 4000\text{Hz}$;

$BW = 1000\text{ Hz}$ for $4000\text{Hz} \leq f(b) < 6000\text{Hz}$; and

$BW = 2000\text{ Hz}$ for $6000\text{Hz} \leq f(b) < 8000\text{Hz}$.

Also, an estimation of the spectrum N_L relating to background noise is performed in a background noise spectrum estimation module 24 by using the magnitude or power spectrum S , in which the steps for the estimation of the spectrum N_L relating to background noise include a number of calculation steps as represented in a block diagram shown in Figure 3.

Firstly in a leak-frequency calculation module 302, a value Leakfrequency $E1$ according to known techniques is calculated from the magnitude or power spectrum S so that the frequency of each FFT bin leaks or spreads to a preset number, preferably two, of neighbouring FFT bins where $E1$ is the maximum magnitude or power spectrum S value within this range.

10

The result E1 from the leak-frequency module 302 is then used in a Freqmax calculation module 304 in which the estimation of the spectrum relating to background noise continues using equation (2), which is:

$$E2(b) = \max_{i=1}^{nav} [E1(i)], \text{ for } i=b, \dots, b+nav, 0 \leq b \leq N \quad (2)$$

in which $nav = \begin{cases} 0 & \text{for } f(b) < 1000\text{Hz} \\ BW / B1 & \text{for } f(b) \geq 1000\text{Hz} \end{cases}$ and $E1(N) = E1(i), \text{ for } i > N;$

where:

E2(b) is the output of the Freqmax module 304;

b, i is the FFT bin number;

f(b) is the frequency of FFT bin b;

B1 is the width of the FFT bin;

and preferably

BW=150 Hz for $1000\text{Hz} \leq f(b) < 1500\text{Hz};$

BW=250 Hz for $1500\text{Hz} \leq f(b) < 2000\text{Hz};$

BW=350 Hz for $2000\text{Hz} \leq f(b) < 3000\text{Hz};$

BW=500 Hz for $3000\text{Hz} \leq f(b) < 4000\text{Hz};$

BW=1000 Hz for $4000\text{Hz} \leq f(b) < 6000\text{Hz};$ and

BW=2000 Hz for $6000\text{Hz} \leq f(b) < 8000\text{Hz}.$

The next step is to find a value RunningMin in a RunningMin calculation module 306, or a local minimum value of the output of the Freqmax module 304. This is done by comparing and selecting the smaller of the output of the Freqmax module 304 obtained in the current calculation frame and the output of the Freqmax module 304 selected in the previous calculation frame, or the smaller of the output of the Freqmax module 304 obtained in the current calculation frame and the maximum value of the output of the Freqmax module 304 obtained during a reference period of m frames known as a phase

clock. This maximum value is preferably limited by the bit-conversion size of the A/D Converter 12. The minimum value E3 according to equation (3) is therefore selected according to:

$$E3(b,j)=\begin{cases} \min[E2(b,j), E2(b,j-1)] & \text{others} \\ \min[E2(b,j), \text{max value}] & \text{at phase clock} \end{cases} \quad (3)$$

The output E3 from the RunningMin module 306 is then saved to a P calculation frame length First-In-First-Out (FIFO) buffer in a FIFO Buffer store module 308 at the beginning every phase clock, in which m is preferably 16 to 32 corresponding to 128 to 256 ms of samples. During this time, the FIFO Buffer module 308 saves preferably 0.5 to 1 sec of data relating to the minimum value E3 to the P calculation frame length FIFO buffer, where P refers to the number of m calculation frames. The preferred P size is 4 so that the P frame length FIFO buffer stores up to 0.5 sec of data in the case when m = 16 calculation frames, and 1 sec of data in the case when m = 32.

During every phase clock or every reference period of m frames, the "best" estimate of the spectrum relating to background noise is obtained from the P calculation frame length FIFO buffer in a MIN of P Calculation Frame select module 310 using the following equation:

$$N_L(b) = \min_{nm=1}^P [E3(b, nm)]$$

where $N_L(b)$ is the estimation of the spectrum relating to background noise as shown in Figure 9; and nm is the order of the calculation frame saved to the FIFO buffer.

After estimation of the spectrums relating to the input noise-contaminated speech signal (S_c) and the background noise ($N_L(b)$) in modules 22 and 24 respectively, a gain vector g is generated in a gain vector calculation module 26 by calculation according to the following equation:

$$g(i) = \left\{ \frac{[Sc(i) - kf * N_L(i)]}{Sc(i)} \right\}^{\frac{1}{a}}, i=0, \dots, N;$$

where kf is a constant factor preferably set between 0.5 to 2, and $a=1$ for fixed-point implementation in Figure 1 and $a=2$ for floating-point implementation in Figure 2. Gain modification of the input noise-contaminated speech signal in a gain vector modification module 28 using the output of the gain vector calculation module 26 involves first the modification of the gain vector g , then using the same to multiply the input noise-contaminated speech signal in the frequency domain derived from the FFT module 18 in the case of the fixed-point device shown in Figure 1, or from an alternative FFT process in the case of the floating-point device shown in Figure 2. Hence, different gain modification processes are appropriately implemented for the different fixed- and floating-point processors, which are described separately hereinafter. Both processes are intended to reduce artifacts and aliasing distortion in the noise-stripped speech signal.

With reference to Figure 4, the floating-point implementation in relation to the gain modification process performed in the gain vector modification module 28 for the floating-point device shown in Figure 1 is described first. Floating-point processors have adequate dynamic range to carry out gain modification processes with very low distortion. The gain vector is transferred back to the time domain by an inverse FFT module, processed using rotating and truncating, then transferred again to the frequency domain by an FFT module. The steps performed in the gain modification process in the respective modules in the gain vector modification module 28 are shown in Figure 4.

A Gmod module 402 is described for setting a minimum gain vector G_{mod} , which includes minimum gain values for the bass, mid-frequency, and high frequency spectral regions. For any minimum gain value $G_{bassmod}$, G_{midmod} , or $G_{highmod}$ where the gain vector g is less than a corresponding preset minimum gain value $G_{bassmin}$, G_{midmin} , or $G_{highmin}$, the respective minimum gain value is set to the predetermined minimum gain value. Preferably, the preset value for $G_{bassmin}$ is 0.15, G_{midmin} is 0.2, and $G_{highmin}$ is 0.15. Otherwise, the minimum gain value follows the gain vector g accordingly:

$$\begin{aligned}
 & \text{13} \\
 G_{\text{bass mod}}(i) &= \begin{cases} G_{\text{bass min}}, g(i) < G_{\text{bass min}} \\ g(i)_{\text{others}} \end{cases} & \text{for } i=0, \dots, 20 \\
 G_{\text{mid mod}}(i) &= \begin{cases} G_{\text{mid min}}, g(i) < G_{\text{mid min}} \\ g(i)_{\text{others}} \end{cases} & \text{for } i=21, \dots, 64 \\
 G_{\text{high mod}}(i) &= \begin{cases} G_{\text{high min}}, g(i) < G_{\text{high min}} \\ g(i)_{\text{others}} \end{cases} & \text{for } i=64, \dots, 128 \\
 G_{\text{mod}} &= [G_{\text{bass mod}} \ G_{\text{mid mod}} \ G_{\text{high mod}}]
 \end{aligned}$$

An IFFT gain module 404 then performs on the minimum gain vector G_{mod} consisting of minimum gain values for the three spectral regions, an $N+1$ complex value Inverse FFT function to yield $2N$ real values in the time domain represented by h_{raw} ,

$$\text{where } h_{\text{raw}} = \text{IFFT}[G_{\text{mod}}]$$

In a Rotate and Truncate module 406, the processes of rotation and truncation, or circular convolution, is performed on h_{raw} by the rotating and truncating h_{raw} , which is the minimum gain vector G_{mod} in the time domain, and saving the rotated and truncated h_{raw} as h_{rot} using

$$h_{\text{rot}}(i) = \begin{cases} h_{\text{raw}}(i + 2 * N - N / 2), i = 0, \dots, N / 2 - 1 \\ h_{\text{raw}}(i - N / 2), i = N / 2, \dots, N - 1 \end{cases}$$

Next in a Window module 408, the rotated and truncated gain vector h_{rot} is processed using a windowing technique, preferably the Hanning windowing technique, to obtain h_{wout} via

$$h_{\text{wout}}(i) = h_{\text{rot}}(i) * w(i), i = 1, \dots, N,$$

where $w(i)$ is a windowing function.

After the windowing operation, an FFT Gain module 410 expands the h_{wout} to $2N$ points as

$$[\text{hwout}, \underbrace{0, \dots, 0}_N],$$

then passes on a $2N$ real value $\text{FFT}[\text{hwout}]$ which is a conversion to the frequency domain.

The gain modification of the input noise-contaminated speech signal is performed through multiplication of the modified gain vector $\text{FFT}[\text{hwout}]$ with the input noise-contaminated speech signal processed by an FFT module 412. The process performed in the FFT module 412 on the input noise-contaminated speech signal is described in greater detail with reference to Figure 2, in which the input noise-contaminated speech signal first passes through a Z^{-N} module 30 for introducing a one-frame delay. In an Expand to $2N$ module 32, N samples of the delayed frame form a frame as X_{in} , and expands the same to $2N$ as

$$[X_{\text{in}}, \underbrace{0, \dots, 0}_N],$$

on which an $\text{FFT}(2)$ module 34 processes for conversion to the frequency domain as X_{fft} as follows:

$$X_{\text{fft}} = \text{FFT}[(X_{\text{in}}, \underbrace{0, \dots, 0}_N)]$$

where X_{in} is N point of input noise-contaminated speech signal.

Then, the X_{fft} is multiplied by the modified gain vector $\text{FFT}[\text{hwout}]$ to produce a noise-stripped speech signal in the frequency domain in a multiplier module 36 as follows:

$$Y = X_{\text{fft}} * \text{FFT}[\text{hwout}]$$

With reference to Figure 5, the gain modification process for the fixed-point implementation is described in greater detail. The gain modification process includes modification of gain vector g and modification of the noise-contaminated input signal represented in frequency domain with the gain vector g . However, modification of the gain vector g only includes setting the minimum for the three bands, followed by mirroring the modified gain vector to $2N$ points.

In a Modification of gain vector module 502, the minimum gain values for the three bands are set accordingly:

$$G_{bass\ mod}(i) = \begin{cases} G_{bass\ min}, g(i) < G_{bass\ min} \\ g(i)_{others} \end{cases} \quad \text{for } i=0, \dots, 20$$

$$G_{mid\ mod}(i) = \begin{cases} G_{mid\ min}, g(i) < G_{mid\ min} \\ g(i)_{others} \end{cases} \quad \text{for } i=21, \dots, 64$$

$$G_{high\ mod}(i) = \begin{cases} G_{high\ min}, g(i) < G_{high\ min} \\ g(i)_{others} \end{cases} \quad \text{for } i=64, \dots, 128$$

$$G_{mod} = [G_{bass\ mod} \ G_{mid\ mod} \ G_{high\ mod}]$$

Next in a Mirror to $2N$ module 504, the minimum gain vector G_{mod} is mirrored to $2N$ points as follows:

$$G_{mod}(i) = G_{mod}(i), \text{ for } i=0, \dots, N; \ G_{mod}(2N-i) = G_{mod}(i), \ i=1, \dots, N-1.$$

The result of mirroring the minimum gain vector G_{mod} is then used to modify the X_{ffts} overlapped FFT of the input noise-contaminated speech signal, in which the X_{ffts} is multiplied with the minimum gain vector G_{mod} in the multiplier module 36 to produce a noise-stripped speech signal as follows:

$$Y = X_{ffts} * G_{mod}$$

In an Inverse Fast Fourier Transform (IFFT) module 38, the treatment of the noise stripped speech signal for both fixed- and floating-point devices proceeds with a $2N$ inverse FFT to convert the noise-stripped signal to the time domain, in which:

$$y_{\text{raw}} = \text{IFFT}[Y],$$

where Y is the noise-stripped speech signal after gain modification in frequency domain, and y_{raw} is the speech signal stripped of the noise in time domain.

The processing then continues with the Overlap yfrain module 40, in which an overlapped noise-stripped signal is generated according to

$$y_{\text{frain}}(i,j) = y_{\text{raw}}(i,j) + y_{\text{raw}}(i+N,j-1), i=0, \dots, N-1$$

A De-emphasis filter 42 utilized only in the fixed-point implementation then processes the overlapped noise-stripped speech signal $y_{\text{frain}}(i,j)$, in which the filter is a first order IIR filter.

A Digital-to-Analog Converter 44 processes the noise-stripped speech signal for conversion back to analog domain for subsequent speech processing applications.

In the foregoing manner, noise-stripping devices according to embodiments of the invention for addressing the foregoing disadvantages of conventional noise-stripping techniques solutions are described. Although only a number of embodiments of the invention are disclosed, it will be apparent to one skilled in the art in view of this disclosure that numerous changes and/or modification can be made without departing from the scope and spirit of the invention.

Claims

1. A method for stripping background noise component from a noise-contaminated speech signal, the method comprising the steps of:
 - digitising the noise-contaminated speech signal to form samples grouped into frames;
 - dividing in the frequency domain the digitised signal into a plurality of frequency bins;
 - storing a plurality of frames of digitised signal equivalent to a preset length of digitised signal in a buffer;
 - estimating the spectrum level of a current frame of digitised signal during a preset period;
 - comparing the spectrum estimate of the current frame of digitised signal with a spectrum estimate representative of an earlier frame of digitised signal and selecting the lower of the two spectrum estimates during the preset period;
 - storing the selected lower spectrum estimate in the buffer during the preset period;
 - assigning the stored and selected lower spectrum estimate as representative of the current frame of digitised signal; and
 - setting as background noise spectrum estimate the minimum value of the stored and selected lower spectrum estimates of the plurality of frames stored in the buffer.
2. The method as in claim 1, wherein the step of storing the plurality of frames includes storing the plurality of frames of digitised signal equivalent to a preset length of at least 0.3 secs of digitised signal in the buffer.
3. The method as in claim 2, wherein the step of storing the plurality of frames includes storing the plurality of frames of digitised signal equivalent to 0.5 to 1 sec of digitised signal in the buffer.
4. The method as in claim 1, wherein the step of estimating the spectrum level includes estimating the spectrum level of the current frame of digitised signal during a preset period of 128 to 256 msecs.

2 March 2005

18

5. The method as in claim 1, wherein the step of comparing the spectrum estimated includes comparing the spectrum estimate of the current frame of digitised signal with a spectrum estimate representative of an earlier adjacent frame of digitised signal.
6. The method as in claim 1, further comprising after the dividing step and before the storing estimate step, the step of adjusting the spectrum level of the frequency divided digitised signal in relation to a frequency bin, the adjustment being dependent on neighbouring frequency bins to which the frequency is leaked.
7. The method as in claim 6, wherein the step of adjusting the spectrum level includes adjusting the spectrum level of the frequency divided digitised signal in relation to a frequency bin exceeding 1 kHz.
8. The method as in claim 7, wherein the step of adjusting the spectrum level includes finding the maximum spectrum value taken between the frequency bin and a next nav number of frequency bins according to

$$E2(b) = \max_{i=1}^{nav} [E1(i)], \text{ for } i=b, \dots, b+nav, 0 \leq b \leq N \quad (2)$$

$$\text{in which } nav = \begin{cases} 0 & \text{for } f(b) < 1000 \text{ Hz} \\ BW / B1 & \text{for } f(b) \geq 1000 \text{ Hz} \end{cases} \quad \text{and} \quad E1(N) = E1(i), \text{ for } i > N;$$

whereby

$E2(b)$ is the maximum spectrum value;

b, i is the frequency bin number;

N is the length of a frame;

$f(b)$ is the frequency of frequency bin b ;

$B1$ is the width of the frequency bin;

$BW=150 \text{ Hz}$ for $1000 \text{ Hz} \leq f(b) < 1500 \text{ Hz}$;

$BW=250 \text{ Hz}$ for $1500 \text{ Hz} \leq f(b) < 2000 \text{ Hz}$;

$BW=350 \text{ Hz}$ for $2000 \text{ Hz} \leq f(b) < 3000 \text{ Hz}$;

BW=500 Hz for $3000\text{Hz} \leq f(b) < 4000\text{Hz}$;
BW=1000 Hz for $4000\text{Hz} \leq f(b) < 6000\text{Hz}$; and
BW=2000 Hz for $6000\text{Hz} \leq f(b) < 8000\text{Hz}$.

9. The method as in claim 1, further comprising the step of multiplying the noise-contaminated speech signal with a gain vector.
10. The method as in claim 9, wherein the step of multiplying the noise-contaminated speech signal with the gain vector includes:
converting the gain vector from frequency to time domain;
performing rotation and truncation operation on the gain vector; and
reforming the rotated and truncated gain vector by inserting zeros and transforming the resultant gain vector to the frequency domain.
11. The method as in claim 9, wherein the step of multiplying the noise-contaminated speech signal with the gain vector includes mirroring the gain vector.
12. The method as in claim 1, further comprising the steps of:
overlapping the plurality of frames; and
performing a windowing operation on the overlapped plurality of frames.
13. A device for stripping background noise component from a noise-contaminated speech signal, the device comprising:
means for digitising the noise-contaminated speech signal to form samples grouped into frames;
means for dividing in the frequency domain the digitised signal into a plurality of frequency bins;
means for storing a plurality of frames of digitised signal equivalent to a preset length of digitised signal in a buffer;
means for estimating the spectrum level of a current frame of digitised signal during a preset period;

means for comparing the spectrum estimate of the current frame of digitised signal with a spectrum estimate representative of an earlier frame of digitised signal and selecting the lower of the two spectrum estimates during the preset period;

means for means for storing the selected lower spectrum estimate in the buffer during the preset period;

means for assigning the stored and selected lower spectrum estimate as representative of the current frame of digitised signal; and

means for setting as background noise spectrum estimate the minimum value of the stored and selected lower spectrum estimates of the plurality of frames stored in the buffer.

14. The device as in claim 13, wherein the means for storing the plurality of frames includes means for storing the plurality of frames of digitised signal equivalent to a preset length of at least 0.3 secs of digitised signal in the buffer.

15. The device as in claim 14, wherein the means for storing the plurality of frames includes means for storing the plurality of frames of digitised signal equivalent to 0.5 to 1 sec of digitised signal in the buffer.

16. The device as in claim 13, wherein the means for estimating the spectrum level includes means for estimating the spectrum level of the current frame of digitised signal during a preset period of 128 to 256 msecs.

17. The device as in claim 13, wherein the means for comparing the spectrum estimated includes means for comparing the spectrum estimate of the current frame of digitised signal with a spectrum estimate representative of an earlier adjacent frame of digitised signal.

18. The device as in claim 13, further comprising means for adjusting the spectrum level of the frequency divided digitised signal in relation to a frequency bin, the adjustment being dependent on neighbouring frequency bins to which the frequency is leaked.

19. The device as in claim 18, wherein the means for adjusting the spectrum level includes means for adjusting the spectrum level of the frequency divided digitised signal in relation to a frequency bin exceeding 1 kHz.

20. The device as in claim 19, wherein the means for adjusting the spectrum level includes means for finding the maximum spectrum value taken between the frequency bin and a next nav number of frequency bins according to

$$E2(b) = \max_{i=1}^{nav} [E1(i)], \text{ for } i=b, \dots, b+nav, 0 \leq b \leq N \quad (2)$$

$$\text{in which } nav = \begin{cases} 0 & \text{for } f(b) < 1000 \text{ Hz} \\ BW / B1 & \text{for } f(b) \geq 1000 \text{ Hz} \end{cases}$$

$$E1(N) = E1(i), \text{ for } i > N;$$

whereby

$E2(b)$ is the maximum spectrum value;

b, i is the frequency bin number;

N is the length of a frame;

$f(b)$ is the frequency of frequency bin b ;

$B1$ is the width of the frequency bin;

$BW = 150 \text{ Hz}$ for $1000 \text{ Hz} \leq f(b) < 1500 \text{ Hz}$;

$BW = 250 \text{ Hz}$ for $1500 \text{ Hz} \leq f(b) < 2000 \text{ Hz}$;

$BW = 350 \text{ Hz}$ for $2000 \text{ Hz} \leq f(b) < 3000 \text{ Hz}$;

$BW = 500 \text{ Hz}$ for $3000 \text{ Hz} \leq f(b) < 4000 \text{ Hz}$;

$BW = 1000 \text{ Hz}$ for $4000 \text{ Hz} \leq f(b) < 6000 \text{ Hz}$; and

$BW = 2000 \text{ Hz}$ for $6000 \text{ Hz} \leq f(b) < 8000 \text{ Hz}$.

21. The device as in claim 13, further comprising means for multiplying the noise-contaminated speech signal with a gain vector.

22. The device as in claim 21, wherein the means for multiplying the noise-contaminated speech signal with the gain vector includes:
- means for converting the gain vector from frequency to time domain;
 - means for performing rotation and truncation operation on the gain vector; and
 - means for reforming the rotated and truncated gain vector by inserting zeros and transforming the resultant gain vector to the frequency domain.
23. The device as in claim 21, wherein the means for multiplying the noise-contaminated speech signal with the gain vector includes means for mirroring the gain vector.
24. The device in claim 13, further comprising:
- means for overlapping the plurality of frames; and
 - means for performing a windowing operation on the overlapped plurality of frames.