



(12) **EUROPEAN PATENT APPLICATION**

(43) Date of publication:
07.05.2008 Bulletin 2008/19

(51) Int Cl.:
G10L 19/06 (2006.01) **G10L 19/02** (2006.01)
G10L 19/08 (2006.01) **G10L 21/02** (2006.01)

(21) Application number: **06022704.8**

(22) Date of filing: **31.10.2006**

(84) Designated Contracting States:
AT BE BG CH CY CZ DE DK EE ES FI FR GB GR HU IE IS IT LI LT LU LV MC NL PL PT RO SE SI SK TR

Designated Extension States:
AL BA HR MK RS

(71) Applicant: **Harman Becker Automotive Systems GmbH**
76307 Karlsbad (DE)

(72) Inventors:
• **Große-Schulte, Dominik**
65232 Traunusstein Hahn (DE)

• **Krini, Mohamed**
89073 Ulm (DE)
• **Schmidt, Gerhard**
89007 Ulm (DE)

(74) Representative: **Grünecker, Kinkeldey, Stockmair & Schwanhäusser**
Anwaltssozietät
Maximilianstrasse 58
80538 München (DE)

(54) **Model-based enhancement of speech signals**

(57) The present invention relates to a method for processing an audio signal comprising the steps of estimating an input-signal-to-noise ratio or a signal-to-noise ratio of the speech input signal, generating an excitation signal based on the speech input signal, extracting a spectral envelope of the speech input signal, generating a reconstructed speech signal on the basis of the excitation signal and the extracted spectral envelope, filtering

the speech input signal by a noise reduction filtering means to obtain a noise reduced signal and combining the reconstructed speech signal and the noise reduced signal on the basis of the speech input-signal-to-noise ratio or the signal-to-noise ratio to obtain an enhance speech output signal. The invention also relates to a signal processing means configured to carry out the herein disclosed method.

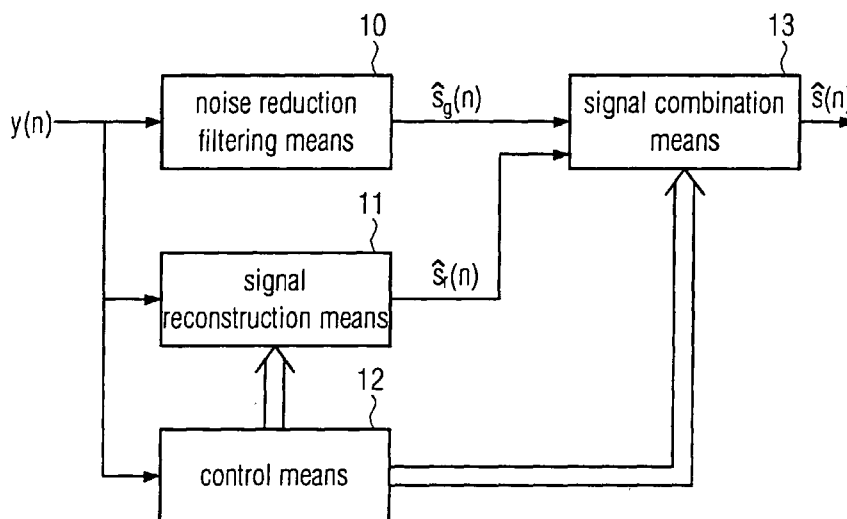


FIG. 2

Description**Field of Invention**

5 [0001] The present invention relates to audio signal processing and, in particular, to the model-based enhancement of the quality of a speech signal by noise reduction and the generation of a reconstructed speech signal by making use of codebooks.

Background of the Invention

10 [0002] One and two-way speech communication of two parties mutually transmitting and receiving audio signals, in particular, speech signals, often suffers from deterioration of the quality of the audio signals by background noise. Background noise in noisy environments can severely affect the quality and intelligibility of voice conversation and can, in the worst case, lead to a complete breakdown of the communication.

15 [0003] One prominent example for speech communication suffering from background noise in noisy environments is hands-free voice communication in vehicles. Some noise reduction must be employed in order to improve the intelligibility of transmitted speech signals. Present vehicle communication systems not only allow for hands-free telephony with remote subscribers at a far end outside the vehicle but also for inter-cabin communication. Microphones and loudspeakers provided for front-seat and back-seat passengers allow for a better acoustical understanding, in particular, if background noises increase during high-speed traveling on motorways.

20 [0004] Another example is given by speech recognition system and speech control systems in a noisy environment. The reliability of the recognition results heavily depend on the contribution of background noise to a detected speech signal and the efficiency of noise suppression means.

25 [0005] In the art, single channel noise reduction methods, e.g., employing spectral subtraction are well-known. These methods, however, are usually limited to (almost) stationary noise perturbations and positive signal-to-noise distances. The processed speech signals are distorted, since according to these methods, perturbations are not eliminated but rather spectral components that are affected by noise are damped. The intelligibility of speech signals is, thus, normally not improved sufficiently. Another method to improve the signal quality in distant talking speech acquisition is the utilization of multi-channel systems, i.e. microphone arrays. However, multi-channel solutions are not only expensive but also cannot guarantee sufficiently enhanced speech signals.

30 [0006] In the some more detail, the enhancement of noisy speech signals is usually performed by means of a time and/or frequency selective attenuation, e.g. by the employment of (modified) Wiener filters, in particular, by means of spectral subtraction. According to the method of spectral subtraction, disturbed parts of the spectrum are suppressed in accordance with a signal-to-noise (SNR) - dependent characteristic. The characteristic and the associated damping factors are usually used for signal processing in the sub-band or spectral domain. Parameters employed in the (Wiener) filter characteristic can, in principle, be stationary or might be time-dependent.

35 [0007] The sub-band signals are usually multiplied with damping factors to obtain estimates for the unperturbed sub-band signals that are combined by means of a synthesis filter bank in order to achieve an enhanced output signal. The damping factors are determined based on estimated power density spectra of the noise (perturbation) and the perturbed audio input signal that is to be enhanced. The power density spectrum of the actual input signal is usually calculated from the squared magnitude of the sub-band signal, i.e. the short-time spectrogram of the sub-band signal. Some damping limit can adaptively be selected.

40 [0008] The known methods for noise reduction in audio and particularly speech signals are, however, not satisfying for a relatively low signal-to-noise ratio. In the art, a signal-to-noise ratio of 10 dB and below results in a significant deterioration of the performance of noise reduction filtering means. In practice, at least some portions of a speech signal are often significantly embedded in background noise.

45 [0009] Thus, despite the recent progress in the art of signal processing, there is still a need for an improved processing of audio, in particular, speech signals, resulting in an enhanced noise reduction and intelligibility of processed speech signals.

50

Description of the invention

[0010] The above mentioned problems are solved by a method for processing a speech input signal according to claim 1, the method comprising the steps of

55

estimating an input-signal-to-noise ratio or a signal-to-noise ratio of the speech input signal;

generating an excitation signal based on the speech input signal;

extracting a spectral envelope of the speech input signal;

generating a reconstructed speech signal on the basis of the excitation signal and the extracted spectral envelope;

5 filtering the speech input signal by a noise reduction filtering means to obtain a noise reduced signal; and

combining the reconstructed speech signal and the noise reduced signal on the basis of the speech input-signal-to-noise ratio or the signal-to-noise ratio to obtain an enhance speech output signal.

10 **[0011]** The input-signal-to-noise ratio (INR) can, e.g., be calculated by the ratio of the short-time spectrogram (squared magnitude of the short-time spectrum) of the speech input signal (which can be a microphone signal generated by a microphone that detects a speech signal from a speaker) and the short-time power density spectrum of noise present in the speech input signal. Methods for the estimation of such spectra are well-known in the art (see, e.g., E. Hänslér, G. Schmidt, "Acoustic Echo and Noise Control", Wiley, Hoboken, NJ, USA, 2004).

15 **[0012]** Alternatively, in all of the herein described examples of the inventive method, the input-signal-to-noise ratio (INR) or the signal-to-noise ratio (SNR) can be used. The latter, e.g., can be calculated from the INR as $SNR(\Omega_{\mu}, n) = \max \{0, INR(\Omega_{\mu}, n) - 1\}$, where n denotes the discrete time index and Ω_{μ} denotes discrete frequency nodes determined by an analysis filter bank or nodes of a discrete Fourier transform used to transform the speech input signal in the frequency domain. In this context, it is noted that the overall signal processing may be performed either in the time or the frequency domain. According to the embodiments described below the speech input signal is transformed into sub-band signals by means of an analysis filter bank, for example.

20 **[0013]** The excitation signal ideally represents the signal that would be detected immediately at the vocal chords of the speaker whose utterance is detected. Discrimination of voiced and unvoiced signal portions is involved in the step of estimating the excitation signal. The excitation signal may be modeled on the basis of the pitch and power of the noisy speech input signal (see description below for further details).

25 **[0014]** The (short-time) spectral envelope is a well-known quantity of particular relevance in speech recognition/synthesis representing the tone color. It may be preferred to employ the robust method of Linear Predictive Coding (LPC) in order to calculate a predictive error filter. The coefficients of the predictive error filter can be used for a parametric determination of the spectral envelope. Alternatively, one may employ models for spectral envelope representation that are based on line spectral frequencies or cepstral coefficients or mel-frequency cepstral coefficients (see also discussion below).

30 **[0015]** According to the present invention a reconstructed speech signal is generated by means of the (short-time) spectral envelope and the estimated excitation signal and this reconstructed speech signal is combined with a noise reduced version of the speech input signal (the employed noise reduction filtering means can be any one known in the art) in dependence on the estimated INR (or SNR, alternatively). In particular, signal portions that exhibit a rather low INR cannot be sufficiently enhanced by simple noise reduction as known in the art. For an SNR of about 10 dB and below, e.g., mere employment of a conventional noise filter may even further deteriorate the noisy speech input signal. For the signal portions with a low INR (SNR) the reconstructed speech signal is used in order to obtain an enhanced speech output signal exhibiting both reduced noise and enhanced intelligibility as compared to the art.

35 **[0016]** The combination of the reconstructed speech signal and the noise reduced signal may be performed in form of a weighted sum of these signals, i.e. each of these signals is multiplied by different weight factors and then the weighted signals are added in order to achieve the enhanced speech output signal. The weights can easily be adapted according to the estimated INR or SNR. According to a specific embodiment, the noise reduced signal and the reconstructed speech signal are each transformed to sub-band signals and multiplied with weights that vary with the discrete frequency nodes determined by an analysis filter bank Ω_{μ} . In a frequency range with an INR below a predetermined threshold the weights are chosen such that the contribution of the reconstructed speech signal to the speech output signal dominates the contribution of the noise reduced signal.

40 **[0017]** Of particular importance for the herein disclosed method for obtaining a speech output signal of enhanced quality from a noisy speech input signal is the estimation of the INR or the SNR and, thus, the estimation of the short-time power density spectrum of the noise (in the following often simply referred to as noise power density spectrum), since the combination of the reconstructed speech signal and the noise reduced signal is controlled depending on the INR or the SNR and, therefore, depending on the estimated noise level.

45 **[0018]** Whereas the INR (SNR) can be estimated by estimating the short-time power density spectrum of noise included in the speech input signal and by determining the short-time spectrogram of the speech input signal, estimating the estimated noise power density spectrum for a discrete time index n may comprise the steps of

55 smoothing the short-time power density spectrum of the speech input signal ($y(n)$) in time to obtain a first smoothed short-time power density spectrum;

smoothing the first smoothed short-time power density spectrum in positive frequency direction to obtain a second smoothed short-time power density spectrum;

5 smoothing the second smoothed short-time power density spectrum in negative frequency direction to obtain a third smoothed short-time power density spectrum; and

determining the minimum of the third smoothed short-time power density spectrum for the discrete time index n and the estimated short-time power density spectrum of noise for a discrete time index $n - 1$.

10 **[0019]** The estimated short-time power density spectrum of noise for a discrete time index $n - 1$ is obtained similarly based on the estimated short-time power density spectrum of noise for a discrete time index $n - 2$, etc.

[0020] By such a particular way of estimating the noise power density spectrum suspension of outliers can be achieved. In order to prevent freezing of the process of estimating the noise power density spectrum it may be preferred to estimate the noise power density spectrum as the maximum of the above-mentioned minimum and a predetermined threshold
15 value. The above-mentioned minimum may be multiplied by a factor of $1 + \epsilon$, where E is a positive real number much less than 1. By an appropriate choice of the value for ϵ , a fast reaction of the estimation to temporal variations can be achieved (see also below).

[0021] In the above-described examples, the excitation signal may be generated by means of an excitation code book. Moreover, the reconstructed speech signal may be generated on the basis of an estimated spectral envelope that is
20 generated from the extracted spectral envelope and by means of a spectral envelope code book. In this case, the spectral code book comprises prototype spectral envelopes trained beforehand which can be used, in particular, for the reconstruction rather noisy portions of the speech input signal.

[0022] From the noisy speech input signal the (voice) pitch of a voiced signal portion can be estimated as known in the art and by means of the estimated pitch period and an excitation code book prepared beforehand the excitation
25 signal can be generated. The excitation code book comprises entries representing weighted sums of sinus oscillations. According to an example the excitation code book can be represented by a matrix C_g of such weighted sums of sinus oscillations wherein the entries in a row $k + 1$ include all of the oscillations of a row k and further include one single additional oscillation. Employment of an excitation code book allows for a reliable and satisfying generation of the voiced part of excitation signals (the unvoiced parts can be generated by noise generators).

[0023] The extracted spectral envelope that is directly obtained (and, e.g., subject to an Infinite Impulse Response
30 smoothing) from the noisy speech input signal represents a reliable estimate for the spectral envelope of the unperturbed speech signal generated by a speaker only for signal portions with a relatively high INR or SNR, respectively. In order to achieve a reliable estimate for the spectral envelope of the unperturbed speech signal in low INR portions a spectral envelope code book is used. The spectral envelope code book is trained beforehand and includes, e.g., 256, trained
35 prototype spectral envelopes. The extracted spectral envelope is compared with the entries of the code book for the signal portions with a relatively high INR. The best matching prototype spectral envelope of the code book is determined.

[0024] According to a specific example, the extracted spectral envelope of the speech input signal is used for the generation of the reconstructed speech signal for signal portions with high INR. For signal portions with low INR the
40 corresponding best matching prototype spectral envelope of the spectral envelope code book is used. Preferably, the resulting spectral envelope used for the generation of the reconstructed speech signal is smoothed in positive and negative frequency direction in order to avoid an abrupt switching from the extracted spectral envelope to the entry of the spectral envelope that best matches the extracted envelope, which may cause perceptible artifacts.

[0025] According to one example of the inventive method employing the above-mentioned spectral envelope code
45 book, a prototype spectral envelope of the spectral envelope code book is determined that best matches the extracted spectral envelope for portions of the speech input signal that exhibit an input-signal-to-noise ratio above a predetermined threshold and the estimated spectral envelope substantially consists, on the one hand, of the determined best matching prototype spectral envelope for the portions of the speech input signal that exhibit an input-signal-to-noise ratio above the predetermined threshold and, on the other hand, of the extracted spectral envelope for the other signal portions.

[0026] In this example the spectral envelope looked-up from the code book is only used for the portion(s) of the speech
50 input signal for which it cannot be expected that the extracted spectral envelope resembles the spectral envelope of the unperturbed speech signal sufficiently to allow for a satisfying reconstructed speech signal in terms of authenticity and intelligibility.

[0027] A particularly efficient way to achieve the above-mentioned combination of the best matching spectral envelope
55 of the code book and the spectral envelope extracted from the speech input signal makes use of a mask function $M(\Omega_\mu, n)$ in the sub-band regime as follows

$$M(\Omega_\mu, n) E_s(e^{j\Omega_\mu}, n) + (1 - M(\Omega_\mu, n)) E_{CB}(e^{j\Omega_\mu}, n)$$

5 where $E_s(e^{j\Omega_\mu}, n)$ and $E_{CB}(e^{j\Omega_\mu}, n)$ denote the smoothed extracted spectral envelope and the best matching spectral envelope of the spectral envelope code book, respectively (for details of the smoothing see detailed description below). The mask function depends on the INR. For example, one might suitably choose $M(\Omega_\mu, n) = 1$, if the INR exceeds a predetermined threshold, and $M(\Omega_\mu, n) = \varepsilon$, where ε is a small positive real number, if the INR lies below the predetermined threshold.

10 **[0028]** Furthermore, it may be preferred that the excitation signal is filtered such that only (mainly) for signal portions for which speech is actually present (detected) an excitation signal (a reconstructed speech signal) is generated. In order to achieve this and, thus, to obtain a further enhancement of the quality of the finally obtained speech output signal, the excitation signal may be generated from filtered excitation sub-band signals $A(e^{j\Omega_\mu}, n)$ that may be generated by transforming the excitation signal obtained as described above into excitation sub-band signals $\tilde{A}(e^{j\Omega_\mu}, n)$. The filtered excitation sub-band signals $A(e^{j\Omega_\mu}, n)$ are obtained by means of a spread noise reduction filtering means $G_s(e^{j\Omega_\mu}, n)$ applied to the unfiltered excitation sub-band signals $\tilde{A}(e^{j\Omega_\mu}, n)$, i.e. $A(e^{j\Omega_\mu}, n) = G_s(e^{j\Omega_\mu}, n) \tilde{A}(e^{j\Omega_\mu}, n)$.

15 **[0029]** A spread conventional noise reduction filter may be used to allow for signal reconstruction in a frequency range showing a low INR (SNR) with filter coefficients

20

$$G_s(e^{j\Omega_\mu}, n) = \max \{ G(e^{j\Omega_\mu}, n), P_0(e^{j\Omega_\mu}, n), P_1(e^{j\Omega_\mu}, n), \dots, P_{M-1}(e^{j\Omega_\mu}, n) \}$$

25 with $P_v(e^{j\Omega_\mu}, n) = G(e^{j\Omega_v}, n) P(e^{j\Omega_\mu - v}, n)$ for $\mu \in \{0, \dots, M-1\}$.

[0030] By $G(e^{j\Omega_\mu}, n)$ conventionally determined damping factors are denoted and $P(e^{j\Omega_\mu}, n)$ denotes a spreading function. For example, a modified Wiener characteristic can be used

30

$$G(e^{j\Omega_\mu}, n) = \max \left\{ G_{\min}(e^{j\Omega_\mu}, n), 1 - \beta(e^{j\Omega_\mu}, n) \frac{\hat{S}_{nn}(\Omega_\mu, n)}{|Y(e^{j\Omega_\mu}, n)|^2} \right\}.$$

35 **[0031]** The same filter characteristic can be used for the above-mentioned noise reduction filtering means. In the context of obtaining filtered excitation sub-band signals, however, a large overestimating factor $\beta(e^{j\Omega_\mu}, n)$ and, in particular, a very high maximum damping, e.g., $G_{\min}(e^{j\Omega_\mu}, n)$ chosen from [0.01, 0.1], are chosen for the employed spread filter.

40 **[0032]** Experiments have shown that it is advantageous to adapt the phases of the sub-band signals of the reconstructed speech signal to the phases of the sub-band signals of the noise reduced signal for signal portions with a relatively high INR (SNR).

45 **[0033]** Thus, according to another example, the generation of the enhanced speech output signal $\hat{S}(n)$ comprises the generation of sub-band signals $(\hat{S}_r(e^{j\Omega_\mu}, n))$ of the reconstructed speech signal $(\hat{S}_r(n))$ and sub-band signals $(\hat{S}_g(e^{j\Omega_\mu}, n))$ of the noise reduced signal $(\hat{S}_g(n))$ and the phases of the sub-band signals $(\hat{S}_r(e^{j\Omega_\mu}, n))$ of the reconstructed speech signal $(\hat{S}_r(n))$ are adapted to the phases of the sub-band signals $(\hat{S}_g(e^{j\Omega_\mu}, n))$ of the noise reduced signal $(\hat{S}_g(n))$ in dependence on the input-signal-to-noise ratio ($INR(\Omega_\mu, n)$), in particular, by

$$\hat{S}_{r,mod}(e^{j\Omega_\mu}, n) = \begin{cases} \frac{\hat{S}_r(e^{j\Omega_\mu}, n)}{\hat{S}_g(e^{j\Omega_\mu}, n)} \hat{S}_g(e^{j\Omega_\mu}, n), & \text{if } INR(\Omega_\mu, n) > \text{some threshold} \\ \hat{S}_r(e^{j\Omega_\mu}, n), & \text{else.} \end{cases}$$

55

[0034] It is noted that the above-mentioned sub-band signals $\hat{S}_g(e^{j\Omega_\mu}, n)$ and $\hat{S}_r(e^{j\Omega_\mu}, n)$ are either generated by analysis filter banks from the noise reduced full band signal output by the noise reduction filtering means and the reconstructed full band speech signal or they are directly output by the noise reduction filtering means and by the process of reconstructing the speech signal.

[0035] The present invention also provides a computer program product, comprising one or more computer readable media having computer-executable instructions for performing the steps of the method according to one of the above described examples.

[0036] The above-mentioned problems of the art are also solved by a signal processing means for processing a speech input signal according to claim 12, comprising

a noise reduction filtering means configured to receive the speech input signal and to output a noise reduced signal;

a signal reconstruction means configured to receive the speech input signal and to extract the spectral envelope from the speech input signal and to generate an excitation signal on the basis of the speech input signal and to generate and output a reconstructed speech signal based on the extracted spectral envelope and the excitation signal;

a signal combination means configured to receive the noise reduced signal and the reconstructed speech signal and to output an enhanced speech output signal as a combination of the noise reduced signal and the reconstructed speech signal; and

a control means configured to receive the speech input signal and to control the signal reconstruction means and the signal combination means based on the input-signal-to-noise ratio or the signal-to-noise ratio of the speech input signal determined by the noise reduction filtering means or the control means.

[0037] The signal processing means may further comprise at least one analysis filter bank configured to transform the speech input signal in speech input sub-band signals and at least one synthesis filter bank configured to synthesize sub-band signals output by the signal combination means or by the noise reduction filtering means and the signal reconstruction means.

[0038] The signal combination means may generate the speech output signal by building a weighted sum of the noise reduced signal and the reconstructed speech signal. The input-signal-to-noise ratio or the signal-to-noise ratio may be estimated by the control means of the noise reduction filtering means by estimating the short-time power density spectrum of noise included in the speech input signal and determining the short-time spectrogram of the speech input signal.

[0039] According to an embodiment the signal processing means comprises

an excitation code book database;

a spectral envelope code book database;

an excitation estimator configured to generate the excitation signal by means of the excitation code book database;

a spectral envelope estimator configured to generate an estimated spectral envelope by means of the spectral envelope code book database; and

the signal reconstruction means is configured to generate and output the reconstructed speech signal based on the estimated spectral envelope and the excitation signal.

[0040] The excitation estimator and the spectral envelope estimator can be configured to generate the excitation signal and the estimated spectral envelope, respectively, as discussed above in the context of the inventive method.

[0041] The control means of the above mentioned examples of the signal processing means can be configured to determine the input-signal-to-noise ratio or the signal-to-noise ratio of the speech input signal and to de-activate the signal reconstruction means, if it is determined that the input-signal-to-noise ratio or the signal-to-noise ratio of the speech input signal exceeds a predetermined threshold for the entire speech input signal. Thus, the relatively expensive reconstruction of the speech signal can be saved, if the perturbation of the input speech signal is not so severe that one has to expect that the noise reduction filtering means might fail in sufficiently enhancing the quality of the speech input signal.

[0042] The present invention also provides a hands-free telephony system, comprising the signal processing means according to one of the above-described examples. It is also provided herein speech recognition means comprising the

signal processing means according to one of the above-described examples.

[0043] Moreover, it is provided a speech dialog system or voice control system comprising the above-mentioned speech recognition means or the signal processing means according to one of the above-described examples.

[0044] Since the quality and, thereby, in particular, the intelligibility of the achieved speech signal that was subject to the inventive signal processing can be enhanced significantly, the above mentioned systems represent examples of preferred applications. Telephone conversation and speech recognition can particularly benefit from the herein disclosed processing of speech signals, since the signal portions of the speech input signal that are heavily perturbed can be replaced by a reconstructed speech signal.

[0045] Additional features and advantages of the invention will be described with reference to the drawings:

Figure 1 is a flow chart illustrating the principle of an example of the herein disclosed method comprising combining a noise reduced speech signal and a reconstructed speech signal.

Figure 2 is block diagram illustrating basic components of an example of the herein disclosed signal processing means comprising a signal reconstruction means and a signal combination means.

Figure 3 shows an example of the signal reconstruction means of Figure 2 comprising a spectral envelope code book and an excitation code book used for the generation of a reconstructed speech signal.

Figure 4 shows an example of a spreading function used for the generation of an excitation signal.

[0046] Basic steps of an example of the inventive method are illustrated in Figure 1. A noisy speech signal is input 1, e.g., by means of one or more microphones, in a signal processing means. A portion of the speech input signal is detected 2 that shows a signal-to-noise ratio (SNR) below a predetermined threshold. This portion of the signal is heavily perturbed by background noise and, thus, conventional noise reduction would fail.

[0047] According to the present example, the spectral envelope is extracted from the speech input signal and the corresponding spectral envelope of the unperturbed speech signal is estimated 3 (details are given below) from the signal portions that are not heavily perturbed by noise. Further, an excitation signal is estimated 4. The process of estimating the excitation signal comprises an excitation classification of the speech input signal in voiced and unvoiced parts. Based on the estimated spectral envelope and the estimated excitation signal the speech input signal is reconstructed 5 as it is known in the art.

[0048] On the other hand, the speech input signal is simply noise reduced 5 by some known noise reduction filtering means. Alternatively, only the portion of the input signal that is not heavily affected by background noise is filtered by the noise reduction filtering means. Finally, the noise reduced signal (portion) and the reconstructed speech signal are combined 7 to obtain a speech output signal with an enhanced quality. Combination of the noise reduced signal and the reconstructed speech signal can be carried out in form of a weighted summation with the weights being dependent on the actual SNR.

[0049] Figure 2 shows basic components of an example of the inventive signal processing means. The signal processing means comprises a conventional noise reduction filtering means 10. A speech input signal $y(n)$ is filtered by this means and a noise reduced signal $\hat{s}_g(n)$ is output. By the argument n the discrete time index is denoted.

[0050] The speech input signal $y(n)$ is also received by a signal reconstruction means 11 and a control means 12. The signal reconstruction means 11 is configured to reconstruct a speech signal on the basis of a feature analysis of the speech input signal $y(n)$. In particular, it is configured to estimate a spectral envelope of the unperturbed speech signal on the basis of the extracted spectral envelope of the speech input signal $y(n)$ and making use of a spectral envelope code book comprising prototype spectral envelopes trained beforehand. The signal reconstruction means 11 is also configured to estimate an unperturbed excitation signal by means of an excitation code book (for details see description below).

[0051] The control means 12 according to the present example is configured to estimate the short-time power density spectrum of the noise and to detect the short-time spectrogram of the speech input signal $y(n)$. For this, the input signal $y(n)$ has to be subject of a short-time Fourier transform or filtering by an analysis filter bank. The SNR or the input-signal-to-noise ratio (INR) of the speech input signal $y(n)$ is also estimated by the control means 12. The control means 12 is also configured to perform a voiced-unvoiced classification of the speech input signal $y(n)$. The operation of the signal reconstruction means 11 is controlled by the results obtained by the control means 12.

[0052] Moreover, the control means 12 controls a signal combination means 13 receiving the reconstructed speech signal $\hat{s}_r(n)$ output by the signal reconstruction means 11 as well as the noise reduced signal $\hat{s}_g(n)$ output by the noise reduction filtering means 10. The control means 12 controls the signal combination means 13 to mix the reconstructed speech signal $\hat{s}_r(n)$ and the noise reduced signal $\hat{s}_g(n)$ or portions of these signals according to the noise level in different signal portions. Signal portions that are heavily perturbed by noise are replaced by the corresponding portions of the

reconstructed speech signal. The signal $\hat{s}(n)$ output by the signal combination means 13 represents an enhanced speech signal.

[0053] According to an example the configuration shown in Figure 2 includes an analysis filter bank that analyzes the speech input signal $y(n)$ into sub-band signals or short-time spectra $Y(e^{j\Omega_\mu}, n)$ with the frequency nodes Ω_μ ($\mu = 0, 1, \dots, M-1$) as it is known in the art. In this case the components 10, 11 and 12 of Figure 2 receive the sub-band signals $Y(e^{j\Omega_\mu}, n)$ and operate in the frequency domain. In this case, the output signal $\hat{s}(n)$ is eventually obtained by means of a synthesis filter bank. It is noted that, in principle, the overall signal processing can be performed either in the time or the frequency domain.

[0054] In the following, the components of the signal processing means shown in Figure 2 are described by way of example in more detail.

[0055] An accurate estimation of the noise that is present in the speech input signal $y(n)$ is crucial for the successful operation of the present invention. This estimate can be performed by the noise reduction filtering means 10 and/or the control means 12. According to an embodiment, it is assumed that speech pauses are present in the perturbed speech input signal $y(n)$ during which the noise can be directly measured. Thus, the short-time spectrogram of $y(n)$, i.e. $|Y(e^{j\Omega_\mu}, n)|^2$ can, in principle, be determined in speech pauses and used for the estimate for the short-time power density spectrum of the background noise in a straightforward way.

[0056] However, according to a preferred example the process of estimating the short-time power density spectrum of the noise that is present in the speech input signal comprises smoothing of the short-time power density spectrum of the speech input signal in time and in frequency as well as a minimum search. Smoothing in time of the short-time power density spectrum of the speech input signal can be performed as an Infinite Impulse Response (IIR) smoothing according to

$$\bar{S}_{yy}(\Omega_\mu, n) = \lambda_T \bar{S}_{yy}(\Omega_\mu, n-1) + (1 - \lambda_T) |Y(e^{j\Omega_\mu}, n)|^2$$

with $0 \leq \lambda_T < 1$. The smaller λ_T is chosen the faster the reaction of the estimation is.

[0057] An IIR smoothing in frequency can, e.g., be performed as

$$\bar{S}'_{yy}(\Omega_\mu, n) = \begin{cases} \bar{S}_{yy}(\Omega_\mu, n), & \text{if } \mu = 0 \\ \lambda_F \bar{S}'_{yy}(\Omega_{\mu-1}, n) + (1 - \lambda_F) \bar{S}_{yy}(\Omega_\mu, n), & \text{if } \mu \in \{1, \dots, M-1\} \end{cases}$$

followed by

$$\bar{S}''_{yy}(\Omega_\mu, n) = \begin{cases} \bar{S}'_{yy}(\Omega_\mu, n), & \text{if } \mu = M-1 \\ \lambda_F \bar{S}''_{yy}(\Omega_{\mu+1}, n) + (1 - \lambda_F) \bar{S}'_{yy}(\Omega_\mu, n), & \text{if } \mu \in \{0, \dots, M-2\} \end{cases}$$

with $0 \leq \lambda_F < 1$. This smoothing in frequency avoids the occurrence of outliers that could result in perceptible artifacts in the processed signal. The estimated short-time power density spectrum of the noise that is used for the subsequent signal processing can be calculated as follows

$$\hat{S}_{nn}(\Omega_\mu, n) = \max \left\{ S_{nn, \min}, \min \left\{ \hat{S}_{nn}(\Omega_\mu, n-1), \bar{S}''_{yy}(\Omega_\mu, n) \right\} (1 + \varepsilon) \right\}$$

with $0 < \varepsilon \ll 1$ and the limiting threshold $S_{nn, \min}$ that guarantees that the estimated short-time power density spectrum is never zero. By the parameter ε the reaction speech of the estimation can be selected. It is to be chosen greater than zero in order to guarantee that reaction to a temporal increase of the noise power density is possible.

[0058] Based on the short-time power density spectrum of the noise $\hat{S}_{nn}(\Omega_\mu, n)$ the control unit 12 can estimate the input-signal-to-noise ratio (INR) by $INR(\Omega_\mu, n) = |Y(e^{j\Omega_\mu}, n)|^2 / \hat{S}_{nn}(\Omega_\mu, n)$. The INR is used for several steps of the subsequent signal processing. In particular, the combination of the reconstructed speech signal $\hat{s}_r(n)$ and the noise reduced signal $\hat{s}_g(n)$ depends on the actual INR. Alternatively, the signal-to-noise ratio (SNR) that is given by $SNR(\Omega_\mu, n) = \max \{0, INR(\Omega_\mu, n) - 1\}$ can be used rather than the INR.

[0059] The control means 12 also performs voiced/unvoiced classification of the speech input signal $y(n)$. According to one example, a sound portion present in the speech input signal $y(n)$ is classified as voiced (unvoiced), if a classification parameter $t_c(n)$ ($0 \leq t_c(n) \leq 1$) is large (small). The classification parameter $t_c(n)$ results from a non-linear mapping of the quantity $r_{INR}(n)$ that is given by $r_{INR}(n) = INR_{high}(n) / (INR_{low}(n) + \Delta_{INR})$ where the constant Δ_{INR} prevents division by zero and

$$INR_{high}(n) = \frac{1}{\mu_3 - \mu_2 + 1} \sum_{\mu=\mu_2}^{\mu_3} INR(\Omega_\mu, n)$$

$$INR_{low}(n) = \frac{1}{\mu_1 - \mu_0 + 1} \sum_{\mu=\mu_0}^{\mu_1} INR(\Omega_\mu, n)$$

where the normalized frequencies Ω_{μ_0} , Ω_{μ_1} , Ω_{μ_2} and Ω_{μ_3} are chosen to correspond to the real frequencies 300 Hz, 1050 Hz, 3800 Hz and 5200 Hz, respectively, as it is appropriate in the context of speech signal processing. In the simplest case a binary classification can be obtained by $t_c(n) = f(r_{INR}(n)) = 1$, if $r_{INR}(n)$ is below a suitable threshold, and zero else.

[0060] Note that the explained example makes use of the fact that unvoiced portions of the speech input signal $y(n)$ show the dominant power density in the high frequency range whereas voiced portions show the dominant power density in the low frequency range.

[0061] Figure 3 shows in some detail the signal reconstruction means 11 of Figure 2 according to an advantageous example. An analysis filter bank 21 is used to obtain sub-band signals $Y(e^{j\Omega_\mu}, n)$ from the speech input signal $y(n)$. Based on the sub-band signals $Y(e^{j\Omega_\mu}, n)$ the excitation signal spectrum $A(e^{j\Omega_\mu}, n)$ and the spectral envelope $E(e^{j\Omega_\mu}, n)$ of the unperturbed speech signal (the speech signal without any noise contribution) are estimated by the signal reconstruction means 11 to obtain a spectrum of the reconstructed speech signal $\hat{S}_r(e^{j\Omega_\mu}, n) = A(e^{j\Omega_\mu}, n) E(e^{j\Omega_\mu}, n)$. The reconstructed speech signal $\hat{s}_r(n)$ of the total band can be obtained from the spectrum $\hat{S}_r(e^{j\Omega_\mu}, n)$ by means of a synthesis filter bank 22.

[0062] It is noted that, alternatively, the spectrum $\hat{S}_r(e^{j\Omega_\mu}, n)$ might be output and combined with a corresponding spectrum output by the noise reduction filtering means $\hat{S}_g(e^{j\Omega_\mu}, n)$. In other word, depending on the specific embodiment the signal combination means 13 of Figure 2 may be configured to combine sub-band signals received from the noise reduction filtering means 10 and the signal reconstruction means 11 and subsequently synthesize the combined sub-band signals or to receive total band signals from these means 10 and 11.

[0063] The estimation of the spectral envelope of the unperturbed speech signal is achieved as follows. First, the spectral envelope $E_s(e^{j\Omega_\mu}, n)$ of the speech input signal $y(n)$ is extracted by a spectral envelope estimator 23 employing a method as known in the art. Linear Prediction Coding (LPC) or a spectral analysis may be used to extract the spectral envelope. According to one example the spectral envelope is directly estimated by a double IIR smoothing method according to

$$E_s(e^{j\Omega_\mu}, n) = \begin{cases} \tilde{E}_s(e^{j\Omega_\mu}, n), & \text{if } \mu = M-1 \\ \lambda_E E_s(e^{j\Omega_{\mu+1}}, n) + (1 - \lambda_E) \tilde{E}_s(e^{j\Omega_\mu}, n), & \text{if } \mu \in \{0, \dots, M-2\} \end{cases}$$

with

$$\tilde{E}_S(e^{j\Omega_\mu}, n) = \begin{cases} |Y(e^{j\Omega_\mu}, n)|, & \text{if } \mu = 0 \\ \lambda_E \tilde{E}_S(e^{j\Omega_{\mu-1}}, n) + (1 - \lambda_E) |Y(e^{j\Omega_\mu}, n)|, & \text{if } \mu \in \{1, \dots, M-1\} \end{cases}$$

with a smoothing constant chosen from $0 \leq \lambda_E < 1$. A suitable choice, e.g., is $\lambda_E = 0.5$.

[0064] The thus extracted spectral envelope is a satisfying approximation for the spectral envelope of the unperturbed speech signal only for signal portions that are not significantly affected by noise. In order to achieve a reliable spectral envelope even for signal portions with a low INR (SNR), a spectral envelope codebook 24 is employed that has been trained beforehand. The spectral envelope codebook 24 includes logarithmic representations of prototype spectral envelopes of particular sounds $E_{CB, \log}(e^{j\Omega_\mu}, 0)$ to $E_{CB, \log}(e^{j\Omega_\mu}, N_{CB, e} - 1)$. A suitable code book may have a size of $N_{CB, e} = 256$.

[0065] The spectral envelope codebook 24 is searched for the entry that best matches the extracted spectral envelope $E_S(e^{j\Omega_\mu}, n)$ for signal portions with a high INR. For this a normalized logarithmic version of the extracted spectral envelope is calculated $\tilde{E}_{S, \log}(e^{j\Omega_\mu}, n) = 20 \log_{10} E_S(e^{j\Omega_\mu}, n) - E_{S, \log, \text{norm}}(n)$ with

$$E_{S, \log, \text{norm}}(n) = \frac{\sum_{\mu=0}^{M-1} M(\Omega_\mu, n) 20 \log_{10} E_S(e^{j\Omega_\mu}, n)}{\sum_{\mu=0}^{M-1} M(\Omega_\mu, n)}$$

where the mask function $M(\Omega_\mu, n)$ depends on the INR, e.g., by $M(\Omega_\mu, n) = g(\text{INR}(\Omega_\mu, n))$, where the mapping function g maps the values of the INR to the interval $[0, 1]$. Resulting values close to one indicate a low noise level, i.e. a good SNR or INR, respectively. In the simplest case one might choose a binary function g that maps to 1, if the INR exceeds a predetermined threshold, e.g., 2 or 4, and to some small but finite real value (to avoid division by zero in the above equation), if the INR falls below that threshold.

[0066] The spectral envelopes of the spectral code book 24 are normalized accordingly and search for the best matching entry of the code book is performed on the basis of a logarithmic INR weighted magnitude distance

$$m_{\text{opt}}(n) = \arg \min_m \sum_{\mu=0}^{M-1} M(\Omega_\mu, n) \left| \tilde{E}_{S, \log}(e^{j\Omega_\mu}, n) - \tilde{E}_{CB, \log}(e^{j\Omega_\mu}, n, m) \right|$$

$$\text{with } \tilde{E}_{CB, \log}(e^{j\Omega_\mu}, n, m) = E_{CB, \log}(e^{j\Omega_\mu}, m) - E_{CB, \log, \text{norm}}(n, m) \quad (m = 0, \dots, N_{cb, e})$$

and

$$E_{CB, \log, \text{norm}}(n, m) = \frac{\sum_{\mu=0}^{M-1} M(\Omega_\mu, n) E_{CB, \log}(e^{j\Omega_\mu}, m)}{\sum_{\mu=0}^{M-1} M(\Omega_\mu, n)}$$

[0067] By argmin the argument of the minimum function is denoted that returns the value of m for which

$\sum_{\mu=0}^{M-1} M(\Omega_{\mu}, n) \left| \tilde{E}_{S, \log}(e^{j\Omega_{\mu}}, n) - \tilde{E}_{CB, \log}(e^{j\Omega_{\mu}}, n, m) \right|$ assumes a minimum. The thus obtained spectral envelope of the spectral envelope code book 24 is linearized and re-normalized

$$E_{CB}(e^{j\Omega_{\mu}}, n) = 10^{(\tilde{E}_{CB, \log}(e^{j\Omega_{\mu}}, n, m_{opt}(n)) + E_{S, \log, norm}(n))/20}$$

[0068] For the reconstruction of the speech signal the obtained spectral envelope $E_{CB}(e^{j\Omega_{\mu}}, n)$ can be used for the portion of the speech input signal with a low INR (SNR) and the extracted spectral envelope $E_S(e^{j\Omega_{\mu}}, n)$ for portions with a high INR (SNR). According to a specific example, the following spectral envelope is output by the spectral envelope estimator 23

$$E(e^{j\Omega_{\mu}}, n) = \begin{cases} \tilde{E}(e^{j\Omega_{\mu}}, n), & \text{if } \mu = M-1 \\ \lambda_{mix} E(e^{j\Omega_{\mu+1}}, n) + (1 - \lambda_{mix}) \tilde{E}(e^{j\Omega_{\mu}}, n), & \text{if } \mu \in \{0, \dots, M-2\} \end{cases}$$

with

$$\tilde{E}(e^{j\Omega_{\mu}}, n) = \begin{cases} \tilde{E}(e^{j\Omega_{\mu}}, n), & \text{if } \mu = 0 \\ \lambda_{mix} \tilde{E}(e^{j\Omega_{\mu-1}}, n) + (1 - \lambda_{mix}) \tilde{E}(e^{j\Omega_{\mu}}, n), & \text{if } \mu \in \{1, \dots, M-1\} \end{cases}$$

and

$\tilde{E}(e^{j\Omega_{\mu}}, n) = M(\Omega_{\mu}, n) E_S(e^{j\Omega_{\mu}}, n) + (1 - M(\Omega_{\mu}, n)) E_{CB}(e^{j\Omega_{\mu}}, n)$. The smoothing constant is chosen between 0 and 1 and may, e.g., be chosen as $\lambda_{mix} = 0.3$.

[0069] The signal reconstruction means 11 also comprises an excitation estimator 25 making use of an excitation code book 26 for estimating an excitation signal that can be shaped with the spectral envelope $E(e^{j\Omega_{\mu}}, n)$ output by the spectral envelope estimator 23 in order to obtain a reconstructed speech signal.

[0070] In order to estimate an excitation signal $a(n)$, first, voiced and unvoiced signal portions are determined as mentioned above. Unvoiced portions $\tilde{a}_u(n)$ of an excitation signal $\tilde{a}(n)$ are generated by a noise generator. In order to achieve a voiced portion $\tilde{a}_v(n)$ of an excitation signal $\tilde{a}(n)$ the (voice) pitch has to be determined. This can be done by any method known in the art (see, e.g., W. Hess, "Pitch Determination of Speech Signals", Springer Berlin, 1983). Based on the determined pitch a voiced portion $\tilde{a}_v(n)$ of an excitation signal $\tilde{a}(n)$ is generated by means of the excitation code book 26 including entries that represent weighted sums of sinus oscillations, e.g.,

$$c_{s,k}(l) = \sum_{m=0}^k 0.99^m \sin\left(\frac{2\pi l(m+1)}{L}\right)$$

where L denotes the length of each code book entry. The entries $C_{s,k}(l)$ are coefficients of a matrix \mathbf{C}_a read out for the generation of the voiced portion $\tilde{a}_v(n)$ of an excitation signal $\tilde{a}(n)$ as $\tilde{a}_v(n) = C_{s,l_z(n)}(l_s(n))$ where $l_z(n)$ denotes the index of the row and $l_s(n)$ the index of the column of the matrix \mathbf{C}_a formed by the coefficients $C_{s,k}(l)$.

[0071] According to the present example, the index of the row is calculated as

$$l_z(n) = \text{round} \left(\frac{T_0 (\text{round}(n/r))}{2} - 1 \right)$$

5

where T_0 is the period of the pitch (that, in general, depends on time) and r/n represents the down-sampled calculation of the period of the pitch (the pitch is calculated every r sampling instants of the speech input signal). The round operation returns the next smaller or larger integer of the argument.

10 **[0072]** The index of the column is calculated by $l_s(n) = \text{round}(\tilde{l}_s(n))$, with

$$15 \quad \tilde{l}_s(n) = \begin{cases} \tilde{l}_s(n-1) + \Delta_s(n), & \text{if } \tilde{l}_s(n-1) + \Delta_s(n) < L - 1.5 \\ \tilde{l}_s(n-1) + \Delta_s(n) - L, & \text{else} \end{cases}$$

with the increment $\Delta_s(n) = L / (T_0 (\text{round}(n/r)))$. The subtraction by 1.5 in the above equation is provided such that the index of the column satisfies the relation $0 \leq l_s(n) \leq L-1$.

20 **[0073]** The excitation signal $\tilde{a}(n)$ is calculated by a weighted summation of the voiced portion $\tilde{a}_v(n)$ and the unvoiced one $\tilde{a}_u(n)$. According to one example, the excitation signal $\tilde{a}(n)$ is calculated by

$$25 \quad \tilde{a}(n) = t_c(\text{round}(n/r)) \tilde{a}_v(n) + [1 - t_c(\text{round}(n/r))] \tilde{a}_u(n).$$

[0074] This signal is input in an analysis filter bank to generate sub-band signals $\tilde{A}(e^{j\Omega_\mu}, n)$ that are subsequently weighted by a noise reduction filter with filter coefficients $G_s(e^{j\Omega_\mu}, n)$ in order to generate an excitation signal only for signal portions in which speech was detected. A spread conventional filter is used for the noise reduction filter to allow for signal reconstruction in frequency ranges showing a low INR (SNR).

30 **[0075]** According to one example the filter coefficients are determined for $\mu \in \{0, \dots, M-1\}$ by

35

$$G_s(e^{j\Omega_\mu}, n) = \max \{ G(e^{j\Omega_\mu}, n), P_0(e^{j\Omega_\mu}, n), P_1(e^{j\Omega_\mu}, n), \dots, P_{M-1}(e^{j\Omega_\mu}, n) \}$$

40 with $P_v(e^{j\Omega_\mu}, n) = G(e^{j\Omega_v}, n) P(e^{j\Omega_{\mu-v}}, n)$.

[0076] By $G(e^{j\Omega_\mu}, n)$ conventionally determined damping factors (filter coefficients) are denoted and $P(e^{j\Omega_m}, n)$ denotes a spreading function. An example of the frequency response of a suitable real-valued positive spreading function is shown in Figure 4. The damping factors, for example, may be determined according to a modified Wiener characteristic including the short-time power density spectrum of the noise and the spectrogram of the speech input signal $y(n)$ can be used

45

$$50 \quad G(e^{j\Omega_\mu}, n) = \max \left\{ G_{\min}(e^{j\Omega_\mu}, n), 1 - \beta(e^{j\Omega_\mu}, n) \frac{\hat{S}_{nn}(\Omega_\mu, n)}{|Y(e^{j\Omega_\mu}, n)|^2} \right\}.$$

55

[0077] The same characteristic can be used in the noise reduction filtering means 10 of Figure 2. In the present context, however, a large overestimating factor $\beta(e^{j\Omega_\mu}, n)$ and, in particular, a very high maximum damping, e.g., $G_{\min}(e^{j\Omega_\mu}, n) = 0.01$ or 0.1 , are chosen for the spread filter. The excitation estimator 25 according to the present example

outputs the estimated excitation spectrum $A(e^{j\Omega_\mu, n}) = G_s(e^{j\Omega_\nu, n}) \tilde{A}(e^{j\Omega_\mu - \nu, n})$. The synthesis filter bank 22 converts the spectrum of the reconstructed speech signal $\hat{S}(e^{j\Omega_\mu, n}) = A(e^{j\Omega_\nu, n}) E(e^{j\Omega_\mu - \nu, n})$ into the total band reconstructed speech signal $\hat{S}_r(n)$ that is output by the signal reconstruction means 11 (see also Figure 2).

[0078] The signal combination means 13 of Figure 2 is configured to combine the reconstructed speech signal $\hat{S}_r(n)$ and the noise reduced signal $\hat{S}_g(n)$ received from the conventional noise reduction filtering means 10. The combination is performed, e.g., in the sub-band regime. Thus, in the present example, the reconstructed speech signal $\hat{S}_r(n)$ and the noise reduced signal $\hat{S}_g(n)$ have to be analyzed into sub-band signals $\hat{S}_r(e^{j\Omega_\mu, n})$ and $\hat{S}_g(e^{j\Omega_\mu, n})$, respectively, by means of analyzing filter banks. Alternative, the noise reduction filtering means 10 could output sub-band signals $\hat{S}_g(e^{j\Omega_\mu, n})$ and the signal reconstruction means 11 could output sub-band signals $\hat{S}_r(e^{j\Omega_\mu, n})$ that then could be directly used by the signal combination means 13.

[0079] Before combining the sub-band signals $\hat{S}_r(e^{j\Omega_\mu, n})$ and $\hat{S}_g(e^{j\Omega_\mu, n})$, it is preferred to adapt the phases of the reconstructed speech signal to the noisy speech input signal (or the output signal of the conventional noise reduction filtering means 10, which usually only influences the amplitude and not the phase of the signal that is to be enhanced). According to an example this can be achieved by

$$\hat{S}_{r, \text{mod}}(e^{j\Omega_\mu, n}) = \begin{cases} \left| \frac{\hat{S}_r(e^{j\Omega_\mu, n})}{\hat{S}_g(e^{j\Omega_\mu, n})} \right| \hat{S}_g(e^{j\Omega_\mu, n}), & \text{if } \text{INR}(\Omega_\mu, n) > \text{some threshold} \\ \hat{S}_r(e^{j\Omega_\mu, n}), & \text{else.} \end{cases}$$

[0080] The modified sub-band signal $\hat{S}_{r, \text{mod}}(e^{j\Omega_\mu, n})$ and the noise reduced sub-band signal $\hat{S}_g(e^{j\Omega_\mu, n})$ are subject to a weighted summation resulting in signals

$$\hat{S}(e^{j\Omega_\mu, n}) = H_g(e^{j\Omega_\mu, n}) \hat{S}_g(e^{j\Omega_\mu, n}) + H_r(e^{j\Omega_\mu, n}) \hat{S}_{r, \text{mod}}(e^{j\Omega_\mu, n}),$$

which are transformed by a synthesis filter bank comprised in the signal combination means 13 in order to obtain the enhanced speech output signal $\hat{S}(n)$. The weights $H_g(e^{j\Omega_\mu, n})$ and $H_r(e^{j\Omega_\mu, n})$ depend on the INR and are determined by the control means 12. According to a preferred example, the weights can be determined by mean values of the INR obtained by the help of ρ Mel filters, $\rho \in \{0, 1, \dots, M_{\text{mel}} - 1\}$, with frequency responses $F_\rho(e^{j\Omega_\mu})$. For a sampling rate of 11025 Hz, e.g., $M_{\text{mel}} = 16$ can suitably be chosen. The averaged INR can be calculated by

$$\text{INR}_{\text{av}}(\rho, n) = \frac{\sum_{\mu=0}^{M-1} F_\rho(e^{j\Omega_\mu}) \text{INR}(\Omega_\mu, n)}{\sum_{\mu=0}^{M-1} F_\rho(e^{j\Omega_\mu})}.$$

[0081] According to a specific embodiment, the weights $H_g(e^{j\Omega_\mu, n})$ and $H_r(e^{j\Omega_\mu, n})$ can be determined from $\text{INR}_{\text{av}}(\rho, n)$ by means of a binary characteristic $f_{\text{mix}}(\text{INR}_{\text{av}}(\rho, n)) = 1$, if $\text{INR}_{\text{av}}(\rho, n) > \text{some threshold}$, e.g., chosen from the interval [4, 10], and $f_{\text{mix}}(\text{INR}_{\text{av}}(\rho, n)) = 0$, else. It is noted that other non-binary characteristics can also be applied. By means of the above-mentioned or any other chosen characteristic $f_{\text{mix}}(\text{INR}_{\text{av}}(\rho, n))$ the weights for the combination of the modified sub-band signal $\hat{S}_{r, \text{mod}}(e^{j\Omega_\mu, n})$ and the noise reduced sub-band signal $\hat{S}_g(e^{j\Omega_\mu, n})$ are calculated by

$$H_g(e^{j\Omega_\nu}, n) = \sum_{\rho=0}^{M_{\text{mel}}} f_{\text{mix}}(\text{INR}_{\text{av}}(\rho, n)) F_\rho(e^{j\Omega_\nu})$$

5

and

10

$$H_r(e^{j\Omega_\nu}, n) = 1 - H_g(e^{j\Omega_\nu}, n).$$

15

[0082] All previously discussed embodiments are not intended as limitations but serve as examples illustrating features and advantages of the invention. It is to be understood that some or all of the above described features can also be combined in different ways.

20

Claims

1. Method for processing a speech input signal ($y(n)$), comprising

25

estimating an input-signal-to-noise ratio (INR) or a signal-to-noise ratio (SNR) of the speech input signal ($y(n)$);
 generating an excitation signal based on the speech input signal ($y(n)$);
 extracting a spectral envelope of the speech input signal ($y(n)$);
 generating a reconstructed speech signal ($\hat{s}_r(n)$) on the basis of the excitation signal and the extracted spectral envelope;

30

filtering the speech input signal ($y(n)$) by a noise reduction filtering means (10) to obtain a noise reduced signal ($\hat{s}_g(n)$); and
 combining the reconstructed speech signal ($\hat{s}_r(n)$) and the noise reduced signal ($\hat{s}_g(n)$) on the basis of the speech input-signal-to-noise ratio (INR) or the signal-to-noise ratio (SNR) to obtain an enhanced speech output signal ($\hat{s}(n)$).

35

2. The method according to claim 1, further comprising

40

weightening the reconstructed speech signal ($\hat{s}_r(n)$) based on the speech input-signal-to-noise ratio (INR) or the signal-to-noise ratio (SNR) to obtain a weighted reconstructed speech signal;
 weightening the noise reduced signal ($\hat{s}_g(n)$) based on the speech input-signal-to-noise ratio (INR) or the signal-to-noise ratio (SNR) to obtain a weighted noise reduced signal; and wherein

45

the step of combining the reconstructed speech signal ($\hat{s}_r(n)$) and the noise reduced signal ($\hat{s}_g(n)$) consists of adding the weighted reconstructed speech signal and the weighted noise reduced signal to obtain the enhanced speech output signal ($\hat{s}(n)$).

3. The method according to claim 1 or 2, wherein the step of estimating the speech input-signal-to-noise ratio (INR) or the signal-to-noise ratio (SNR) comprises the steps of

50

estimating the short-time power density spectrum of noise included in the speech input signal ($y(n)$); and
 determining the short-time spectrogram of the speech input signal ($y(n)$).

4. The method according to claim 3, wherein estimating the short-time power density spectrum of noise for a discrete time index n from a short-time power density spectrum of noise for a discrete time index $n-1$ comprises the steps of

55

smoothing the short-time power density spectrum of the speech input signal ($y(n)$) in time to obtain a first smoothed short-time power density spectrum;
 smoothing the first smoothed short-time power density spectrum in positive frequency direction to obtain a

second smoothed short-time power density spectrum;
 smoothing the second smoothed short-time power density spectrum in negative frequency direction to obtain
 a third smoothed short-time power density spectrum; and
 determining the minimum of the third smoothed short-time power density spectrum for the discrete time index
 n and the estimated short-time power density spectrum of noise for a discrete time index n - 1.

- 5 5. The method according to one of the preceding claims, wherein the excitation signal is generated by means of an excitation code book.
- 10 6. The method according to one of the preceding claims, wherein the reconstructed speech signal ($\hat{s}_r(n)$) is generated on the basis of an estimated spectral envelope that is generated from the extracted spectral envelope and by means of a spectral envelope code book.
- 15 7. The method according to claim 6, wherein a prototype spectral envelope of the spectral envelope code book is determined that best matches the extracted spectral envelope for portions of the speech input signal that exhibit an input-signal-to-noise ratio above a predetermined threshold and wherein the estimated spectral envelope substantially consist of the determined best matching prototype spectral envelope for the portions of the speech input signal that exhibit an input-signal-to-noise ratio above the predetermined threshold and of the extracted spectral envelope for the other signal portions.
- 20 8. The method according to claim 7, wherein the estimated spectral envelope is calculated in the sub-band regime from the weighted sum of the extracted spectral envelope smoothed in frequency $E_s(e^{j\Omega_\mu}, n)$ and the determined best matching prototype spectral envelope $E_{CB}(e^{j\Omega_\mu}, n)$

$$M(\Omega_\mu, n) E_s(e^{j\Omega_\mu}, n) + (1 - M(\Omega_\mu, n)) E_{CB}(e^{j\Omega_\mu}, n)$$

where the mask function $M(\Omega_\mu, n)$ depends on the input-signal-to-noise ratio.

- 30 9. The method according to one of the claims 6 to 8, wherein the excitation signal is generated from filtered excitation sub-band signals ($A(e^{j\Omega_\mu}, n)$), and wherein the filtered excitation sub-band signals ($A(e^{j\Omega_\mu}, n)$) are obtained by means of a spread noise reduction filtering means.
- 35 10. The method according to one of the preceding claims, wherein the generation of the enhanced speech output signal ($\hat{s}(n)$) comprises the generation of sub-band signals ($\hat{S}_r(e^{j\Omega_\mu}, n)$) of the reconstructed speech signal ($\hat{s}_r(n)$) and sub-band signals ($\hat{S}_g(e^{j\Omega_\mu}, n)$) of the noise reduced signal ($\hat{s}_g(n)$) and the phases of the sub-band signals ($\hat{S}_r(e^{j\Omega_\mu}, n)$) of the reconstructed speech signal ($\hat{s}_r(n)$) are adapted to the phases of the sub-band signals ($\hat{S}_g(e^{j\Omega_\mu}, n)$) of the noise reduced signal ($\hat{s}_g(n)$) in dependence on the input-signal-to-noise ratio ($INR(\Omega_\mu, n)$), in particular, by

$$\hat{S}_{r,mod}(e^{j\Omega_\mu}, n) = \begin{cases} \left| \frac{\hat{S}_r(e^{j\Omega_\mu}, n)}{\hat{S}_g(e^{j\Omega_\mu}, n)} \right| \hat{S}_g(e^{j\Omega_\mu}, n), & \text{if } INR(\Omega_\mu, n) > \text{some threshold} \\ \hat{S}_r(e^{j\Omega_\mu}, n), & \text{else.} \end{cases}$$

- 45 11. Computer program product, comprising one or more computer readable media having computer-executable instructions for performing the steps of the method according to one of the Claims 1 to 10.
- 50 12. Signal processing means for processing a speech input signal ($y(n)$), comprising

a noise reduction filtering means (10) configured to receive the speech input signal ($y(n)$) and to output a noise reduced signal ($\hat{s}_g(n)$);

a signal reconstruction means (11) configured to receive the speech input signal ($y(n)$) and to extract the spectral envelope from the speech input signal ($y(n)$) and to generate an excitation signal on the basis of the speech input signal ($y(n)$) and to generate and output a reconstructed speech signal ($\hat{s}_r(n)$) based on the extracted spectral envelope and the excitation signal;

5 a signal combination means (13) configured to receive the noise reduced signal ($\hat{s}_g(n)$) and the reconstructed speech signal ($\hat{s}_r(n)$) and to output an enhanced speech output signal ($\hat{s}(n)$) as a combination of the noise reduced signal ($\hat{s}_g(n)$) and the reconstructed speech signal ($\hat{s}_r(n)$); and

10 a control means (12) configured to receive the speech input signal ($y(n)$) and to control the signal reconstruction means (11) and the signal combination means (13) based on the input-signal-to-noise ratio or the signal-to-noise ratio of the speech input signal ($y(n)$) determined by the noise reduction filtering means (10) or the control means (12).

13. The signal processing means according to claim 12, further comprising at least one analysis filter bank (21) configured to transform the speech input signal ($y(n)$) in speech input sub-band signals ($Y(e^{j\Omega_\mu}, n)$) and at least one synthesis filter bank (22) configured to synthesize sub-band signals output by the signal combination means (13) or by the noise reduction filtering means (10) and the signal reconstruction means (11).

14. The signal processing means according to claim 12 or 13, wherein the signal reconstruction means (11) comprises

20 an excitation code book database (26);

a spectral envelope code book database (24);

an excitation estimator (25) configured to generate the excitation signal by means of the excitation code book database (26);

25 a spectral envelope estimator (23) configured to generate an estimated spectral envelope by means of the spectral envelope code book database (24); and

wherein the signal reconstruction means (11) is configured to generate and output the reconstructed speech signal ($\hat{s}_r(n)$) based on the estimated spectral envelope and the excitation signal.

15. The signal processing means according to one of the claims 12 to 14, wherein the control means (12) is configured to determine the input-signal-to-noise ratio or the signal-to-noise ratio of the speech input signal ($y(n)$) and to deactivate the signal reconstruction means (11), if it is determined that the input-signal-to-noise ratio or the signal-to-noise ratio of the speech input signal ($y(n)$) exceeds a predetermined threshold for the entire speech input signal ($y(n)$).

16. Hands-free telephony system, comprising the signal processing means according to one of the claims 12 - 15.

17. Speech recognition means comprising the signal processing means according to one of the claims 12 - 15.

18. Speech dialog system or voice control system comprising the speech recognition means according to claim 17 or the signal processing means according to one of the claims 12 - 15.

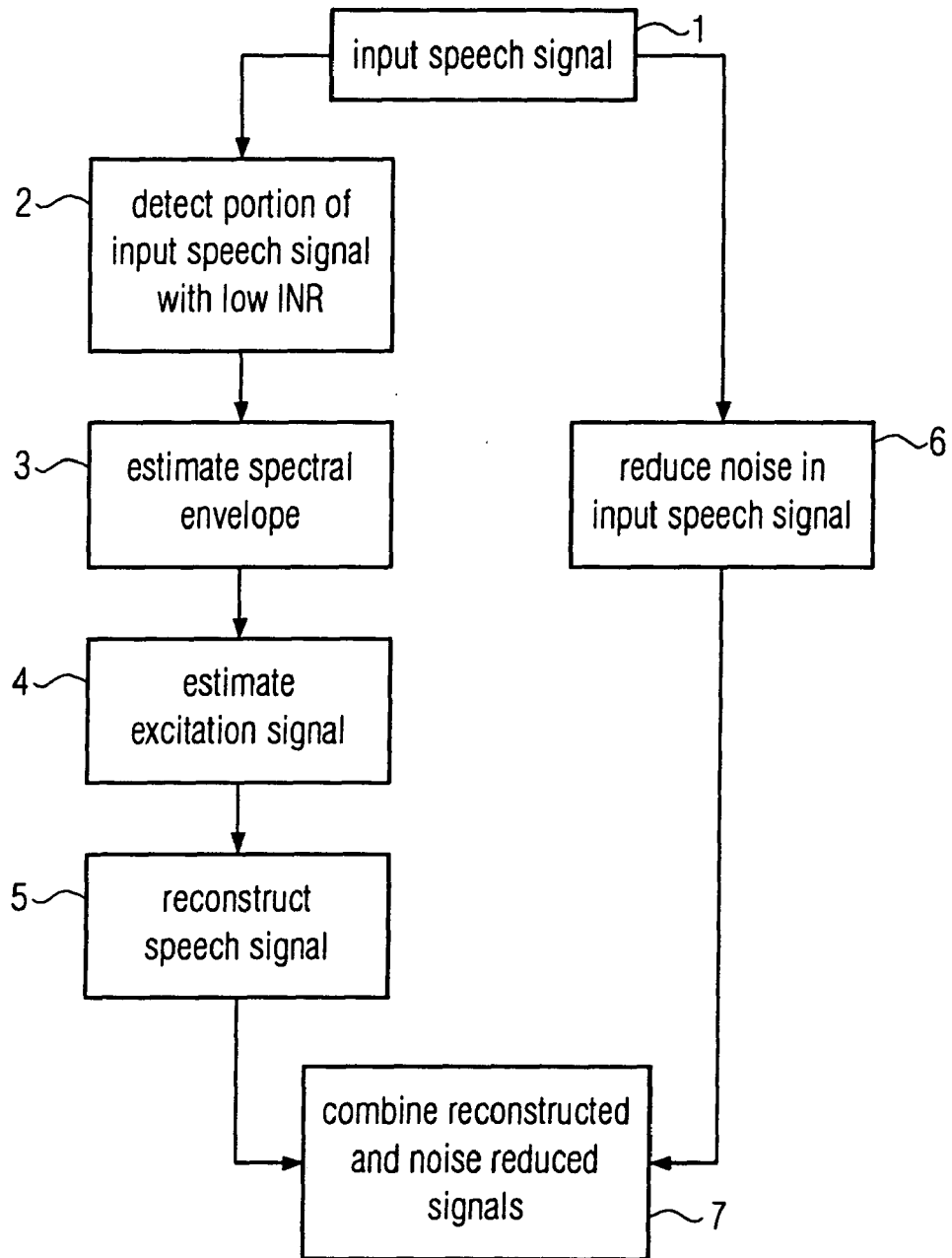


FIG. 1

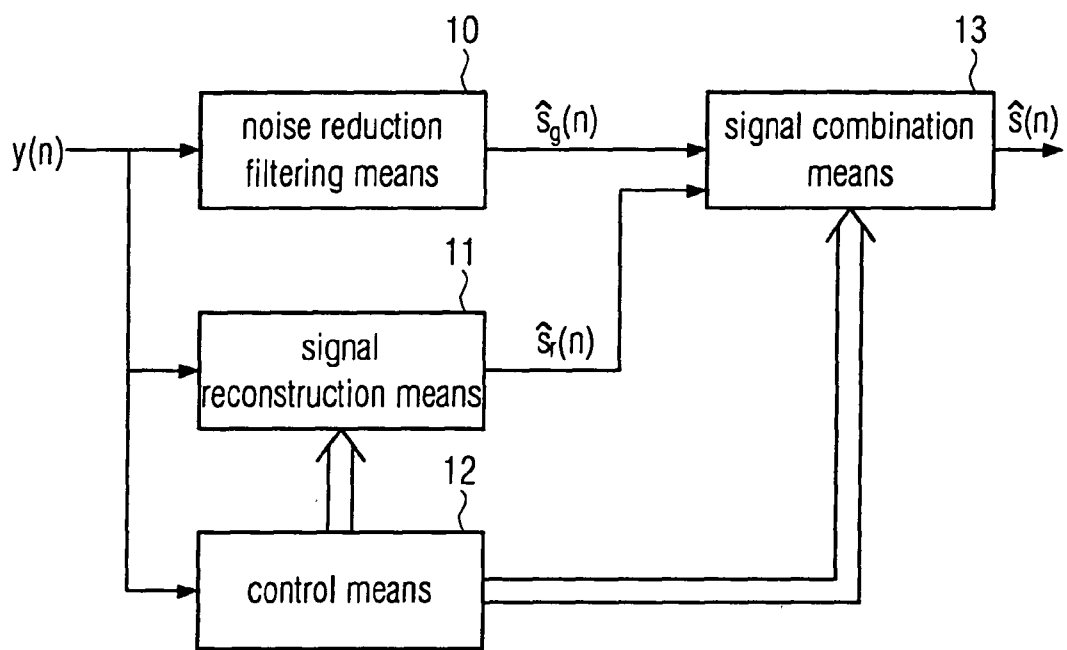


FIG. 2

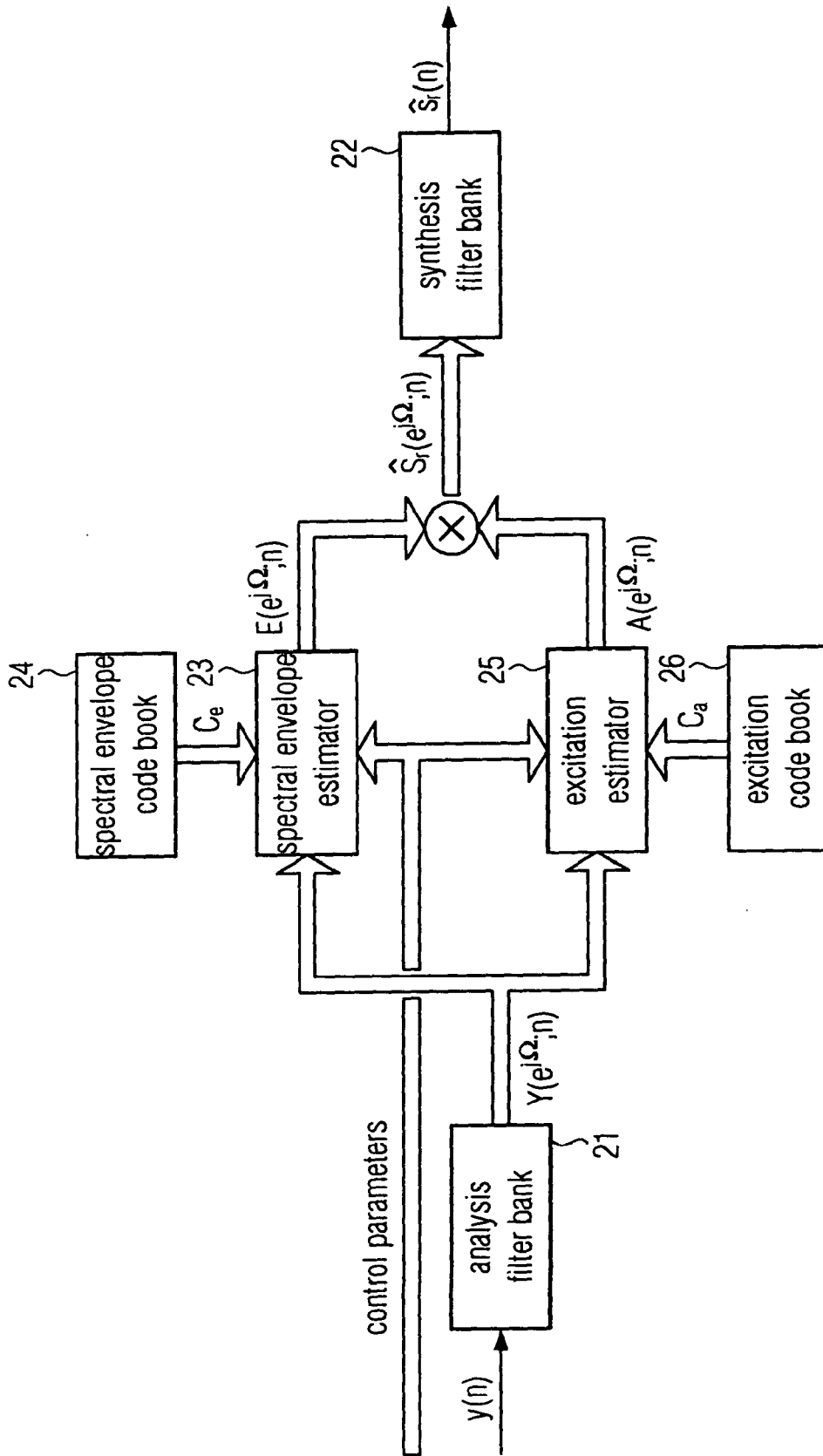


FIG. 3

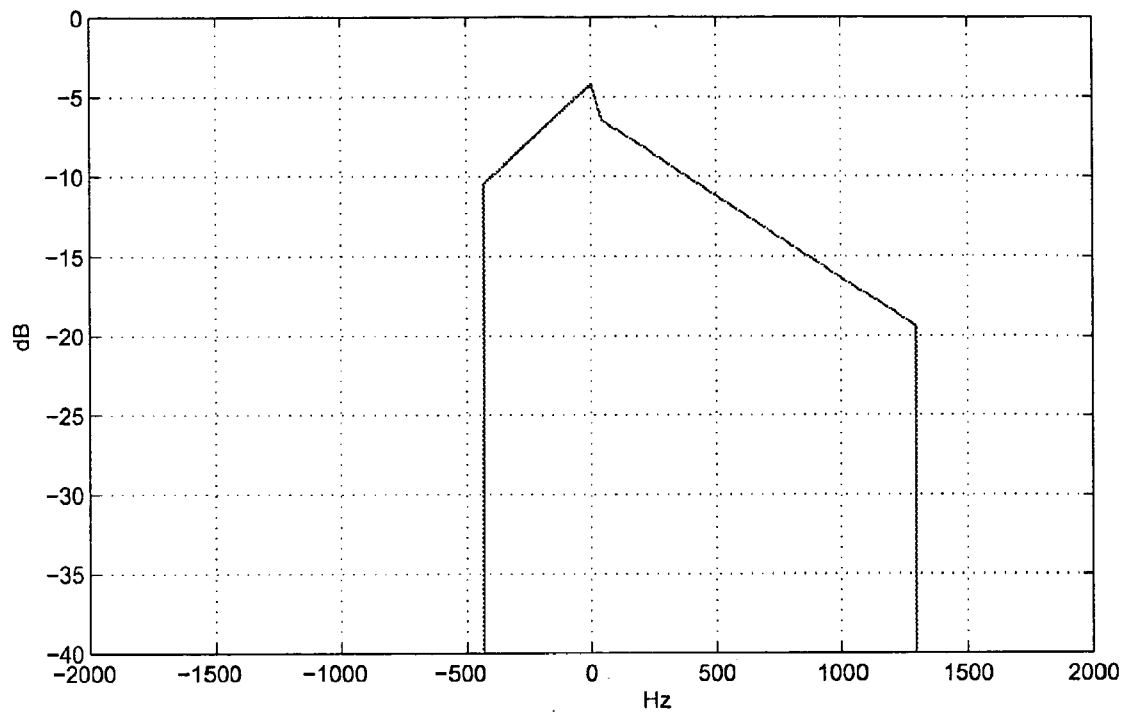


FIG. 4



| DOCUMENTS CONSIDERED TO BE RELEVANT | | | |
|--|---|---|--|
| Category | Citation of document with indication, where appropriate, of relevant passages | Relevant to claim | CLASSIFICATION OF THE APPLICATION (IPC) |
| A | ARSLAN L ET AL: "New methods for adaptive noise suppression" ACOUSTICS, SPEECH, AND SIGNAL PROCESSING, 1995. ICASSP-95., 1995 INTERNATIONAL CONFERENCE ON DETROIT, MI, USA 9-12 MAY 1995, NEW YORK, NY, USA, IEEE, US, vol. 1, 9 May 1995 (1995-05-09), pages 812-815, XP010625357 ISBN: 0-7803-2431-5 * page 1, right-hand column, line 20 - line 36 * * page 2, left-hand column, line 10 - line 15 * * page 2, left-hand column, line 39 - line 46 * * page 2, right-hand column, line 5 - page 3, left-hand column, line 6 * ----- | 1-18 | INV. G10L19/06 G10L19/02 G10L19/08 G10L21/02 |
| A | WO 94/18666 A (BRITISH TELECOMM [GB]; CROZIER PHILIP MARK [GB]; CHEETHAM BARRY MICHAEL) 18 August 1994 (1994-08-18) * page 2, line 6 - line 9 * * page 3, line 6 - line 36 * * page 4, line 6 - line 10 * * page 4, line 19 - line 26 * * page 5, line 20 - line 36 * * page 6, line 31 - page 7, line 3 * ----- | 1-18 | TECHNICAL FIELDS SEARCHED (IPC) G10L |
| A | US 7 065 486 B1 (THYSSEN JES [US]) 20 June 2006 (2006-06-20) * page 1, column 2, line 20 - line 43 * * page 2, column 3, line 41 - line 44 * * page 3, column 5, line 41 - line 45 * * page 3, column 5, line 60 - line 67 * ----- -/-- | 1-18 | |
| 4 The present search report has been drawn up for all claims | | | |
| Place of search Munich | | Date of completion of the search 2 February 2007 | Examiner Aalburg, Stefanie |
| CATEGORY OF CITED DOCUMENTS X : particularly relevant if taken alone Y : particularly relevant if combined with another document of the same category A : technological background O : non-written disclosure P : intermediate document T : theory or principle underlying the invention E : earlier patent document, but published on, or after the filing date D : document cited in the application L : document cited for other reasons & : member of the same patent family, corresponding document | | | |

EPC FORM 1503 03.82 (P04C01)



| DOCUMENTS CONSIDERED TO BE RELEVANT | | | |
|--|---|---|---|
| Category | Citation of document with indication, where appropriate, of relevant passages | Relevant to claim | CLASSIFICATION OF THE APPLICATION (IPC) |
| A | US 5 708 754 A (WYNN WOODSON DALE [US]) 13 January 1998 (1998-01-13) * page 1, column 2, line 17 - line 40 * * page 2, column 4, line 23 - line 27 * * page 2, column 4, line 47 - line 66 * * page 3, column 5, line 9 - line 22 * * page 3, column 5, line 42 - line 52 * * page 3, column 6, line 21 - line 65 * ----- | 1-18 | |
| | | | TECHNICAL FIELDS SEARCHED (IPC) |
| | | | |
| The present search report has been drawn up for all claims | | | |
| Place of search Munich | | Date of completion of the search 2 February 2007 | Examiner Aalburg, Stefanie |
| CATEGORY OF CITED DOCUMENTS X : particularly relevant if taken alone Y : particularly relevant if combined with another document of the same category A : technological background O : non-written disclosure P : intermediate document | | T : theory or principle underlying the invention E : earlier patent document, but published on, or after the filing date D : document cited in the application L : document cited for other reasons & : member of the same patent family, corresponding document | |

4
EPO FORM 1503 03/02 (P04C01)

ANNEX TO THE EUROPEAN SEARCH REPORT
ON EUROPEAN PATENT APPLICATION NO.

EP 06 02 2704

This annex lists the patent family members relating to the patent documents cited in the above-mentioned European search report. The members are as contained in the European Patent Office EDP file on
The European Patent Office is in no way liable for these particulars which are merely given for the purpose of information.

02-02-2007

| Patent document cited in search report | Publication date | Patent family member(s) | Publication date |
|--|------------------|-------------------------|------------------|
| WO 9418666 A | 18-08-1994 | AU 676714 B2 | 20-03-1997 |
| | | AU 6006194 A | 29-08-1994 |
| | | CA 2155832 C | 18-07-2000 |
| | | DE 69420027 D1 | 16-09-1999 |
| | | DE 69420027 T2 | 06-07-2000 |
| | | ES 2137355 T3 | 16-12-1999 |
| | | JP 8506427 T | 09-07-1996 |
| | | NO 953169 A | 11-10-1995 |
| | | SG 49709 A1 | 15-06-1998 |
| | | US 5742927 A | 21-04-1998 |
| | | ----- | ----- |
| US 7065486 B1 | 20-06-2006 | NONE | |
| US 5708754 A | 13-01-1998 | CA 2153170 A1 | 08-06-1995 |
| | | CN 1129486 A | 21-08-1996 |
| | | EP 0681730 A1 | 15-11-1995 |
| | | JP 8506434 T | 09-07-1996 |
| | | PL 310056 A1 | 13-11-1995 |
| | | WO 9515550 A1 | 08-06-1995 |
| | | US 5706394 A | 06-01-1998 |
| | | US 5781883 A | 14-07-1998 |
| ----- | ----- | ----- | ----- |

EPO FORM P0458

For more details about this annex : see Official Journal of the European Patent Office, No. 12/82

REFERENCES CITED IN THE DESCRIPTION

This list of references cited by the applicant is for the reader's convenience only. It does not form part of the European patent document. Even though great care has been taken in compiling the references, errors or omissions cannot be excluded and the EPO disclaims all liability in this regard.

Non-patent literature cited in the description

- **E. HÄNSLER ; G. SCHMIDT.** Acoustic Echo and Noise Control. Wiley, 2004 [0011]
- **W. HESS.** Pitch Determination of Speech Signals. Springer, 1983 [0070]