(19) Europäisches Patentamt
European Patent Office
Office européen des brevets

(11) EP 1 343 145 A1

(12) EUROPEAN PATENT APPLICATION

(43) Date of publication:
10.09.2003 Bulletin 2003/37

(51) Int Cl.[7]: **G10L 19/00**

(21) Application number: **02075973.4**

(22) Date of filing: **08.03.2002**

(84) Designated Contracting States:
AT BE CH CY DE DK ES FI FR GB GR IE IT LI LU MC NL PT SE TR
Designated Extension States:
AL LT LV MK RO SI

(71) Applicant: **Koninklijke KPN N.V.**
9726 AE Groningen (NL)

(72) Inventor: **Beerends, John Gerard**
4585 PG Hengstdijk (NL)

(74) Representative: **Wuyts, Koenraad Maria et al**
Koninklijke KPN N.V.,
Intellectual Property Group,
P.O. BOX 95321
2509 CH The Hague (NL)

(54) Method and system for measuring a sytems's transmission quality

(57) Method and system for measuring the transmission quality of an audio transmission system, an input signal (X) being entered into the system, resulting in an output signal (Y), output by the system, both signals, the input signal and the output signal, being mutually processed, the method comprising that the output signal and/or the input signal of the system under test are scaled in a way that small deviations of the power are compensated, while larger deviations are compensated partially, dependent on the power ratio. A compensation ratio F is calculated from the power representations PX and PY resp. of said input signal (X) and output signal (Y), being equal to the ratio PX/PY. A clipped ratio C is calculated, C being equal to a first clipping value mm for F<mm, or C being equal to a second clipping value MM for F>MM, or otherwise C being equal to F. A softscale ratio S is calculated from a first scaling factor m and a second scaling factor M, with $mm < m \leq 1$ and $MM > M \geq 1$, S being equal to $C^a + C - C(m)^{a-1}$ for $C < m$, 'a' being a first tuning parameter being set to a value > 0 and < 1, or S being equal to $C^a + C - C(M)^{a-1}$ for $C > M$, or otherwise S being equal to C. An artificial reference speech signal may be created, for which the noise levels as present in the original input speech signal are lowered by a scaling factor that depends on the local level of the noise in this input.

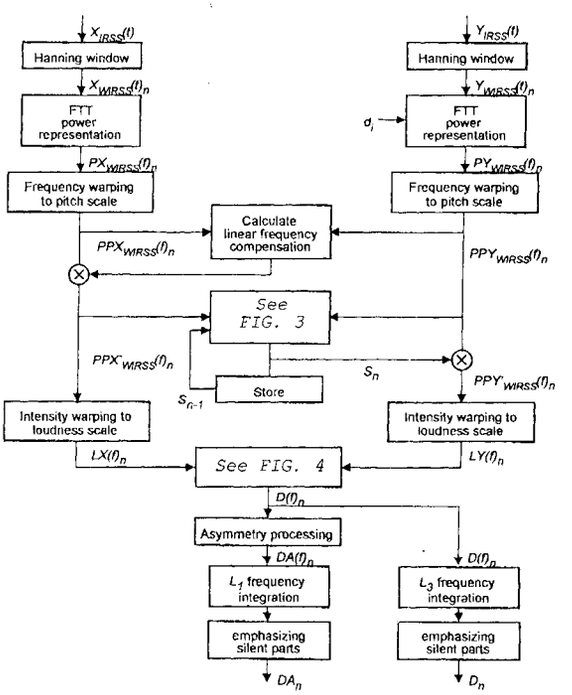

FIG. 2

EP 1 343 145 A1

Description

FIELD OF THE INVENTION

[0001] The invention refers to a method and a system for measuring the transmission quality of a system under test, an input signal entered into the system under test and an output signal resulting from the system under test being processed and mutually compared.

BACKGROUND OF THE INVENTION

[0002] Draft ITU-T recommendation P.862, "Telephone transmission quality, telephone installations, local line networks - Methods for objective and subjective assessment of quality - Perceptual evaluation of speech quality (PESQ), an objective method for en-to-end speech qualtity assessment of narrow-bank telephone networks and speech codecs", ITU-T 02.2001, discloses prior-art PESQ methods and systems.

[0003] Measuring the quality of audio signals, degraded in audio processing or transmission systems, may have poor results for very weak or silent portions in the input signal. The methods and systems known from Recommendation P.862 have the disadvantage that they do not compensate for differences in power level on a frame by frame basis correctly. These differences are caused by gain variations or noise in the input signal. The incorrect compensation leads to low correlations between subjective and objective scores, especially when the original reference input speech signal contains low levels of noise.

[0004] According to a prior-art method and system, disclosed in applicant's EP01200945, improvements are achieved by applying a first scaling step in a pre-processing stage with a first scaling factor which is a function of the reciprocal value of the power of the output signal increased by an adjustment value. A second scaling step is applied with a second scaling factor which is substantially equal to the first scaling factor raised to an exponent having a adjustment value between zero and one. The second scaling step may be carried out on various locations in the device, while the adjustment values are adjusted using test signals with well defined subjective quality scores.

[0005] Both, in the methods and systems of Recommendation P.862 and EP01200945 the degraded output signal is scaled locally to match the reference input signal in the power domain.

[0006] It has been found that the results of the (perceptual) quality measurement process can be improved by application of "soft scaling" at at least one stage of the method and system respectively. Introduction of "soft scaling" instead of "hard scaling" (using "hard" scaling thresholds) is based on the observation and understanding that -the field of the invention relates assessment of audio quality as *experienced* by human users- human audio perception mechanisms rather use "soft thresholds" than "hard thresholds". Based on that observation and a better understanding of how those human audio scaling mechanism works, the present invention presents such "soft scaling" mechanisms, to be added to or inserted into the prior-art method or system respectivley.

SUMMARY OF THE INVENTION

[0007] According to an aspect of the invention the output signal and/or the input signal of a system are scaled, in a way that small deviations of the power are compensated, while larger deviations are compensated partially in a manner that is dependent on the power ratio.

[0008] According to a further elaboration of the invention an artificial reference speech signal may be created, for which the noise levels as present in the original input speech signal are lowered by a scaling factor that depends on the local level of the noise in this input.

[0009] The result of the inventive measures is a more correct prediction of the subjectively perceived end-to-end speech quality for speech signals which contain variations in the local scaling, especially in the case where soft speech parts and silences are degraded by low levels of noise.

[0010] In the softscaling algorithm two different types of signal processing are used to improve the correlation between subjectively perceived quality and objectively measured quality.

[0011] In the first softscale processing, controlled by a first sub-algorithm, the compensation used in Recommendation P.862 to correct for local gain changes in the output signal, is improved by scaling the output (or the input) in such way that small deviations of the power are compensated (preferably per time frame or period) while larger deviations are compensated partially, dependent on the power ratio.

[0012] A preferred simple and effective implementation takes the local powers, i.e. the power in each frame (of e.g. 30 ms.) and calculates a *local compensation ratio F:*

$$F = (PX + \Delta) / (PY + \Delta) *)$$

which F is amplitude clipped at levels mm and MM to get a *clipped ratio C:*

$$C = mm \text{ whenever } F < mm \leq 1.0$$

and

$$C = MM \text{ whenever } F > MM \geq 1.0$$

while otherwise

$$C = F$$

*) "Δ" is used to optimize the value of C for small values of PY.

[0013] The clipped ratio C is then used to calculate a *softscale ratio S* by using factors m and M, with $mm < m \leq 1.0$ and $MM > M \geq 1.0$:

$$S = C^{a} + C - C(m)^{a-1} \text{ whenever } C < m \text{ with } 0.5 < a < 1.0$$

and

$$S = C^{a} + C - C(M)^{a-1} \text{ whenever } C > M \text{ with } 0.5 < a < 1.0$$

while otherwise

$$S = C$$

"a" may be used as a (first) tuning parameter.

[0014] In this way the local scaling in the present invention is equivalent to the scaling as given in the prior-art documents Recommendation P.862 and EP01200945 as long as $m \leq F \leq M$. However for values $F < m$ or $F > M$ the scaling is progressively deviating less from 1.0 then the scaling as given in the prior-art. The softscale factor S is used in the same way F is used in the prior-art methods and systems to compensate the output power in each frame locally.

[0015] In the second softscale processing, controlled by a second sub-algorithm, the compensation used is focussed on low level parts of the input signal.

[0016] When the input signal (reference signal) contains low levels of noise, a transparent speech transport system will give an output speech signal that also contains low levels of noise. The output of the speech transport system is then judged of having *lower* quality then expected on the basis of the noise introduced by the transport system. One would only be aware of the fact that the noise is *not* caused by the transport system if one could listen to the input speech signal and make a comparison. However in most *subjective* speech quality tests the input reference is not presented to the testing subject and consequently the subject judges low noise level differences in the input signal as differences in quality of the speech transport system. In order to have high correlations, in objective test systems, with such subjective tests, this effect has to be emulated in an advanced objective speech quality assessment algorithm.

[0017] The present preferred option of the invention emulates this by effectively creating a new, virtual, artificial reference speech signal in the power representation domain for which the noise power levels are lowered by a scaling factor that depends on the local level of the noise in the input signal. Thus the newly created artificial reference signal converges to zero faster than the original input signal for low levels of this input signal. When the disturbances in the degraded output signal are calculated during low level signal parts, as present in the reference input signal, the difference calculation in the internal representation loudness domain is carried out after scaling of the input loudness signal to a level that goes to zero faster than the loudness of the input signal as it approaches zero.

[0018] According to the prior-art method disclosed in EP01200945, the processing implies mapping of the (degraded) output signal (Y(t)) and the reference signal (X(t)) on *representation signals* LY and LX according to a psycho-physical perception model of the human auditory system. A differential or disturbance signal (D) is determined by "differentiating means" from those representation signals, which disturbance signal is then processed by modelling means in accord-

ance with a cognitive model, in which certain properties of human testees have been modelled, in order to obtain the quality signal Q.

[0019] As said above, the difference calculation in the internal representation loudness domain is, within the scope of the present invention, preferably carried out after scaling the input loudness signal to a level that goes to zero faster than the loudness of the input signal as it approaches zero.

[0020] An effective implementation of this is achieved by using the difference in internal representation in the time-frequency plane calculated from LX(f)n and LY(f)n -see EP01200945- as

$$D(f)n = |LY(f)n - LX(f)n|$$

and replacing this by:

$$D(f)n = |LY(f)n - H(t,f)|$$

with

$$H(t,f) = LX(f)n^{b}/K^{b-1} \text{ for all } LX(f)n < K$$

and

$$H(t,f) = LX(f)n \text{ for all } LX(f)n \geq K$$

[0021] In these formula is b > 1 while K represents the low level noise power criterion per time frequency cell, dependent on the specific implementation.

[0022] This second softscale processing sub-algorithm can also be implemented by replacing the LX(f)n < K criterion by a power criterion in a single time frame i.e.:

$$D(f)n = |LY(f)n - H(t,f)|$$

with

$$H(t,f) = LX(f)n^{b}/K^{b-1} \text{ for all } LX(t) < K'$$

and

$$H(t,f) = LX(f)n \text{ for all } LX(t) \geq K'$$

[0023] In these formula is b > 1 while K' represents the low level noise power criterion per time frame which is dependent on the specific implementation.

BRIEF DESCRIPTION OF THE DRAWINGS

[0024]

Figure 1 shows schematically a prior-art PESQ system, disclosed in ITU-T recommendation P.862.
Figure 2 shows the same PESQ system which, however, is modified to be fit for executing the method as presented above by the use of a first and, preferrably, a second new module.
Figure 3 shows the first new module of the PESQ system.
Figure 4 shows the second new module of the PESQ system.

DETAILED DESCRIPTION OF THE DRAWINGS

[0025] The PESQ system shown in figure 1 compares an original signal (input signal) X(t) with a degraded signal (output signal) Y(t) that is the result of passing X(t) through e.g. a communication system. The output of the PESQ system is a prediction of the perceived quality that would be given to Y(t) by subjects in a subjective listening test.

[0026] In the first step executed by the PESQ system a series of delays between original input and degraded output are computed, one for each time interval for which the delay is significantly different from the previous time interval. For each of these intervals a corresponding start and stop point is calculated. The alignment algorithm is based on the principle of comparing the confidence of having two delays in a certain time interval with the confidence of having a single delay for that interval. The algorithm can handle delay changes both during silences and during active speech parts.

[0027] Based on the set of delays that are found the PESQ system compares the original (input) signal with the aligned degraded output of the device under test using a perceptual model. The key to this process is transformation of both the original and the degraded signals to internal representations (LX, LY), analogous to the psychophysical representation of audio signals in the human auditory system, taking account of perceptual frequency (Bark) and loudness (Sone). This is achieved in several stages: time alignment, level alignment to a calibrated listening level, time-frequency mapping, frequency warping, and compressive loudness scaling.

[0028] The internal representation is processed to take account of effects such as local gain variations and linear filtering that may - if they are not too severe - have little perceptual significance. This is achieved by limiting the amount of compensation and making the compensation lag behind the effect. Thus minor, steady-state differences between original and degraded are compensated. More severe effects, or rapid variations, are only partially compensated so that a residual effect remains and contributes to the overall perceptual disturbance. This allows a small number of quality indicators to be used to model all subjective effects. In the PESQ system, two error parameters are computed in the cognitive model; these are combined to give an objective listening quality MOS (Mean Opinion Score). The basic ideas used in the PESQ system are described in the bibliography references [1] to [5].

The perceptual model in the prior-art PESQ system

[0029] The perceptual model of a PESQ system, shown in figure 1, is used to calculate a distance between the original and degraded speech signal ("PESQ score"). This may be passed through a monotonic function to obtain a prediction of a subjective MOS for a given subjective test. The PESQ score is mapped to a MOS-like scale, a single number in the range of -0.5 to 4.5, although for most cases the output range will be between 1.0 and 4.5, the normal range of MOS values found in an ACR listening quality experiment.

Precomputation of constant settings

[0030] Certain constants values and functions are pre-computed. For those that depend on the sample frequency, versions for both 8 and 16 kHz sample frequency are stored in the program.

FFT window size and sample frequency

[0031] In the PESQ system the time signals are mapped to the time frequency domain using a short term FFT (Fast Fourier Transformation) with a Hann window of size 32 ms. For 8 kHz this amounts to 256 samples per window and for 16 kHz the window counts 512 samples while adjacent frames are overlapped by 50%.

Absolute hearing threshold

[0032] The absolute hearing threshold $P_0(f)$ is interpolated to get the values at the center of the Bark bands that are used. These values are stored in an array and are used in Zwicker's loudness formula.

The power scaling factor

[0033] There is an arbitrary gain constant following the FFT for time-frequency analysis. This constant is computed from a sine wave of a frequency of 1 000 Hz with an amplitude at 29.54 (40 dB SPL) transformed to the frequency domain using the windowed FFT over 32 ms. The (discrete) frequency axis is then converted to a modified Bark scale by binning of FFT bands. The peak amplitude of the spectrum binned to the Bark frequency scale (called the "pitch power density") must then be 10 000 (40 dB SPL). The latter is enforced by a postmultiplication with a constant, the power scaling factor S_P.

The loudness scaling factor

[0034] The same 40 dB SPL reference tone is used to calibrate the psychoacoustic (Sone) loudness scale. After binning to the modified Bark scale, the intensity axis is warped to a loudness scale using Zwicker's law, based on the absolute hearing threshold. The integral of the loudness density over the Bark frequency scale, using a calibration tone at 1 000 Hz and 40 dB SPL, must then yield a value of 1 Sone. The latter is enforced by a postmultiplication with a constant, the loudness scaling factor S_1.

IRS-receive filtering

[0035] As stated in section 10.1.2 it is assumed that the listening tests were carried out using an IRS receive or a modified IRS receive characteristic in the handset. The necessary filtering to the speech signals is already applied in the pre-processing.

Computation of the active speech time interval

[0036] If the original and degraded speech file start or end with large silent intervals, this could influence the computation of certain average distortion values over the files. Therefore, an estimate is made of the silent parts at the beginning and end of these files. The sum of five successive absolute sample values must exceed 500 from the beginning and end of the original speech file in order for that position to be considered as the start or end of the active interval. The interval between this start and end is defined as the active speech time interval. In order to save computation cycles and/or storage size, some computations can be restricted to the active interval.

Short term FFT

[0037] The human ear performs a time-frequency transformation. In the PESQ system this is implemented by a short term FFT with a window size of 32 ms. The overlap between successive time windows (frames) is 50 per cent. The power spectra - the sum of the squared real and squared imaginary parts of the complex FFT components - are stored in separate real valued arrays for the original and degraded signals. Phase information within a single Hann window is discarded in the PESQ system and all calculations are based on only the power representations $PX_{WIRSS}(f)_n$ and $PY_{WIRSS}(f)_n$. The start points of the windows in the degraded signal are shifted over the delay. The time axis of the original speech signal is left as is. If the delay increases, parts of the degraded signal are omitted from the processing, while for decreases in the delay parts are repeated.

Calculation of the pitch power densities

[0038] The Bark scale reflects that at low frequencies, the human hearing system has a finer frequency resolution than at high frequencies. This is implemented by binning FFT bands and summing the corresponding powers of the FFT bands with a normalization of the summed parts. The warping function that maps the frequency scale in Hertz to the pitch scale in Bark does not exactly follow the values given in the literature. The resulting signals are known as the pitch power densities $PPX_{WIRSS}(f)_n$ and $PPY_{WIRSS}(f)_n$.

Partial compensation of the original pitch power density

[0039] To deal with filtering in the system under test, the power spectrum of the original and degraded pitch power densities are averaged over time. This average is calculated over speech active frames only using time-frequency cells whose power is more than 1 000 times the absolute hearing threshold. Per modified Bark bin, a partial compensation factor is calculated from the ratio of the degraded spectrum to the original spectrum. The maximum compensation is never more than 20 dB. The original pitch power density $PPX_{WIRSS}(f)_n$ of each frame n is then multiplied with this partial compensation factor to equalize the original to the degraded signal. This results in an inversely filtered original pitch power density $PPX'_{WIRSS}(f)_n$. This partial compensation is used because severe filtering can be disturbing to the listener. The compensation is carried out on the original signal because the degraded signal is the one that is judged by the subjects in an ACR experiment.

Partial compensation of the distorted pitch power density

[0040] Short-term gain variations are partially compensated by processing the pitch power densities frame by frame. For the original and the degraded pitch power densities, the sum in each frame n of all values that exceed the absolute

hearing threshold is computed. The ratio of the power in the original and the degraded files is calculated and bounded to the range $[3 \cdot 10^{-4}, 5]$. A first order low pass filter (along the time axis) is applied to this ratio. The distorted pitch power density in each frame, n, is then multiplied by this ratio, resulting in the partially gain compensated distorted pitch power density $PPY'_{WIRSS}(f)_n$.

Calculation of the loudness densities

[0041] After partial compensation for filtering and short-term gain variations, the original and degraded pitch power densities are transformed to a Sone loudness scale using Zwicker's law [7].

$$LX(f)_n = S_l \cdot \left(\frac{P_0(f)}{0.5}\right)^{\gamma} \cdot \left[\left(0.5 + 0.5 \cdot \frac{PPX'_{WIRSS}(f)_n}{P_0(f)}\right)^{\gamma} - 1\right]$$

with $P_0(f)$ the absolute threshold and S_1 the loudness scaling factor. Above 4 Bark, the Zwicker power, γ, is 0.23, the value given in the literature. Below 4 Bark, the Zwicker power is increased slightly to account for the so-called recruitment effect. The resulting two-dimensional arrays $LX(f)_n$ and $LY(f)_n$ are called loudness densities.

Calculation of the disturbance density

[0042] The signed difference between the distorted and original loudness density is computed. When this difference is positive, components such as noise have been added. When this difference is negative, components have been omitted from the original signal. This difference array is called the raw disturbance density.

[0043] The minimum of the original and degraded loudness density is computed for each time frequency cell. These minima are multiplied by 0.25. The corresponding two-dimensional array is called the mask array. The following rules are applied in each time-frequency cell:

- If the raw disturbance density is positive and larger than the mask value, the mask value is subtracted from the raw disturbance.
- If the raw disturbance density lies in between plus and minus the magnitude of the mask value the disturbance density is set to zero.
- If the raw disturbance density is more negative than minus the mask value, the mask value is added to the raw disturbance density.

[0044] The net effect is that the raw disturbance densities are pulled towards zero. This represents a dead zone before an actual time frequency cell is perceived as distorted. This models the process of small differences being inaudible in the presence of loud signals (masking) in each time-frequency cell. The result is a disturbance density as a function of time (window number *n)* and frequency, $D(f)_n$.

Cell-wise multiplication with an asymmetry factor

[0045] The asymmetry effect is caused by the fact that when a codec distorts the input signal it will in general be very difficult to introduce a new time-frequency component that integrates with the input signal, and the resulting output signal will thus be decomposed into two different percepts, the input signal and the distortion, leading to clearly audible distortion [2]. When the codec leaves out a time-frequency component the resulting output signal cannot be decomposed in the same way and the distortion is less objectionable. This effect is modelled by calculating an asymmetrical disturbance density $DA(f)_n$ per frame by multiplication of the disturbance density $D(f)_n$ with an asymmetry factor. This asymmetry factor equals the ratio of the distorted and original pitch power densities raised to the power of 1.2. If the asymmetry factor is less than 3 it is set to zero. If it exceeds 12 it is clipped at that value. Thus only those time frequency cells remain, as non-zero values, for which the degraded pitch power density exceeded the original pitch power density.

Aggregation of the disturbance densities

[0046] The disturbance density $D(f)_n$ and asymmetrical disturbance density $DA(f)_n$ are integrated (summed) along the frequency axis using two different *Lp* norms and a weighting on soft frames (having low loudness):

$$D_n = M_n \sqrt[3]{\sum_{f=1,..Number of Barkbands} (\ |D(f)_n|\ W_f\ )^3}$$

$$DA_n = M_n \sum_{f=1,..Number of Barkbands} (\ |DA(f)_n|\ W_f\ )$$

with M_n a multiplication factor, *1/ (power of original frame plus a constant)*$^{0.04}$, resulting in an emphasis of the disturbances that occur during silences in the original speech fragment, and W_f a series of constants proportional to the width of the modified Bark bins. After this multiplication the frame disturbance values are limited to a maximum of 45. These aggregated values, D_n and DA_n, are called frame disturbances.

Zeroing of the frame disturbance

[0047] If the distorted signal contains a decrease in the delay larger than 16 ms (half a window) the repeat strategy as mentioned in 10.2.4 is modified. It was found to be better to ignore the frame disturbances during such events in the computation of the objective speech quality. As a consequence frame disturbances are zeroed when this occurs. The resulting frame disturbances are called D'_n and DA'_n.

Realignment of bad intervals

[0048] Consecutive frames with a frame disturbance above a threshold are called bad intervals. In a minority of cases the objective measure predicts large distortions over a minimum number of bad frames due to incorrect time delays observed by the preprocessing. For those so-called bad intervals a new delay value is estimated by maximizing the cross correlation between the absolute original signal and absolute degraded signal adjusted according to the delays observed by the preprocessing. When the maximimal cross correlation is below a threshold, it is concluded that the interval is matching noise against noise and the interval is no longer called bad, and the processing for that interval is halted. Otherwise, the frame disturbance for the frames during the bad intervals is recomputed and, if it is smaller replaces the original frame disturbance. The result is the final frame disturbances $D''n$ and DA''_n that are used to calculate the perceived quality.

Aggregation of the disturbance within split second intervals

[0049] Next, the frame disturbance values and the asymmetrical frame disturbance values are aggregated over split second intervals of 20 frames (accounting for the overlap of frames: approx. 320 ms) using L_6 norms, a higher p value as in the aggregation over the speech file length. These intervals also overlap 50 per cent and no window function is used.

Aggregation of the disturbance over the duration of the signal

[0050] The split second disturbance values and the asymmetrical split second disturbance values are aggregated over the active interval of the speech files (the corresponding frames) now using L_2 norms. The higher value of p for the aggregation within split second intervals as compared to the lower p value of the aggregation over the speech file is due to the fact that when parts of the split seconds are distorted that split second loses meaning, whereas if a first sentence in a speech file is distorted the quality of other sentences remains intact.

Computation of the PESQ score

[0051] The final PESQ score is a linear combination of the average disturbance value and the average asymmetrical disturbance value. The range of the PESQ score is -0.5 to 4.5, although for most cases the output range will be a listening quality MOS-like score between 1.0 and 4.5, the normal range of MOS values found in an ACR (Absolute Category Rating) experiment.

[0052] Figure 2 is equal to figure 1, with the exception of a first new module, replacing the prior-art module for calculation the local scaling factor and a new second module, replacing the prior-art module for perceptial subtraction.

[0053] The first new module is fit for execution of the method according the invention, comprising means for scaling the output signal and/or the input signal of the system under test, under control of a new, "soft-scaling" algorithm, compensating small deviations of the power, while compensating larger deviations partially, dependent on the power ratio. The first module is depicted in figure 3.

[0054] The second new module is fit for execution of a further elaboration of the invention, comprising means for the creation of an artificial reference speech signal, for which the noise levels as present in the original input speech signal are lowered by a scaling factor that depends on the local level of the noise in this input.

[0055] The operation of both new modules are depicted in the form of flow diagrams, representing the operation of the respective modules. Both modules may be implemented in hardware or in software.

[0056] Figure 3 depicts the operation of the first new module shown in figure 2. The operation of the module in figure 3 is controlled by the first sub-algorithm as represented by the depected flow diagram, improving the compensation function to correct for local gain changes in the output signal, by scaling the output resp. input in such way that small deviations of the power are compensated, preferably per time frame or period, while larger deviations are compensated partially, dependent on the power ratio. The preferred simple and effective implementation of the invention takes the local powers, i.e. the power in each frame (of e.g. 30 ms.) and calculates a *local compensation ratio*

$$F = (PX + \Delta) / (PY + \Delta)$$

Note: PX and PY are the shorter notations of $PPX_{WIRSS}(f)_n$ and $PPY_{WIRSS}(f)_n$ respectively as use in the figures 1, 2 and 3
F is amplitude clipped at levels mm and MM to get a *clipped ratio*

$$C = mm \text{ for } F < mm \leq 1.0 \text{ or } C = MM \text{ for } F > MM \geq 1.0 \text{ or } C = F$$

"Δ" for optimizing C for small values of PX and/or PY)

[0057] The clipped ratio C is used to calculate a *softscale ratio S* by using factors m and M, with $mm < m \leq 1.0$ and $MM > M \geq 1.0$.

[0058] Softscale ratio $S = C^a + C - C(m)^{a-1}$ for $C < m$ $(0.5 < a < 1.0)$ or
$S = C^a + C - C(M)^{a-1}$ for $C > M$ or $S = C$

[0059] In this way the local scaling in the present invention is equivalent to the scaling as given in the prior-art documents Recommendation P.862 and EP01200945 as long as $m \leq F \leq M$. However for values $F < m$ or $F > M$ the scaling is progressively deviating less from 1.0 than the scaling as given in the prior-art. The softscale factor S is used in the same way F is used in the prior-art methods and systems to compensate the output power in each frame locally.

[0060] In the second softscale processing, controlled by a second sub-algorithm, advanced scaling is applied on low level parts of the input signal. When the input signal (reference signal) contains low levels of noise, a transparent speech transport system will give an output speech signal that also contains low levels of noise. The output of the speech transport system is then judged of having *lower* quality then expected on the basis of the noise introduced by the transport system. One would only be aware of the fact that the noise is *not* caused by the transport if system one could listen to the input speech signal and make a comparison. However in most *subjective* speech quality tests the input reference is not presented to the testing subject and consequently the subject judges low noise level differences in the input signal as differences in quality of the speech transport system. In order to have high correlations, in objective test systems, with such subjective tests, this effect has to be emulated in an advanced objective speech quality assessment algorithm. The embodiment of the preferred option of the invention, illustrated in figure 4, emulates this by creating an *artificial* reference speech signal in the power representation domain for which the noise power levels are lowered by a scaling factor that depends on the local level of the noise in the input signal. Thus the artificial reference signal converges to zero faster than the original input signal for low levels of this input signal. When the disturbances in the degraded output signal are calculated during low level signal parts, as present in the reference input signal, the difference calculation in the internal representation loudness domain is carried out after scaling of the input loudness signal to a level that goes to zero faster than the loudness of the input signal as it approaches zero.

[0061] The difference in internal representation in the time-frequency plane is set to

$$D(f)n = |LY(f)n - LX(f)n^b/K^{b-1}| \text{ for } LX(f)n < K$$

or

$$D(f)n = |LY(f)n - LX(f)n| \text{ for } LX(f)n \geq K.$$

[0062] In these formula is b > 1 while K represents the low level noise power criterion per time frequency cell.

[0063] As an alternative the second softscale processing sub-algorithm can also be implemented by replacing the LX(f)n < K criterion by a power criterion in a single time frame. In this alternative option the difference in internal representation in the time-frequency plane is set to

$$D(f)n = |LY(f)n - LX(f)n^{b}/K^{b-1}| \text{ for } LX(t) < K'$$

or

$$D(f)n = |LY(f)n - LX(f)n| \text{ for } LX(t) \geq K'.$$

[0064] In these alternative formula is b > 1 while K' represents the low level noise power criterion per time frame.

References incorporated herein by references

[0065]

[1] BEERENDS (J.G.), STEMERDINK (J.A.): A Perceptual Speech-Quality Measure Based on a Psychoacoustic Sound Representation, J. *Audio Eng. Soc.,* Vol. 42, No. 3, pp. 115-123, March 1994.

[2] BEERENDS (J.G.): Modelling Cognitive Effects that Play a Role in the Perception of Speech Quality, *Speech Quality Assessment,* Workshop papers, Bochum, pp. 1-9, November 1994.

[3] BEERENDS (J.G.): Measuring the quality of speech and music codecs, an integrated psychoacoustic approach, 98th AES Convention, pre-print No. 3945, 1995.

[4] HOLLIER (M.P.), HAWKSFORD (M.O.), GUARD (D.R.): Error activity and error entropy as a measure of psychoacoustic significance in the perceptual domain, IEE Proceedings - Vision, Image and Signal Processing, 141 (3), 203-208, June 1994.

[5] RIX (A.W.), REYNOLDS (R.), HOLLIER (M.P.): Perceptual measurement of end-to-end speech quality over audio and packet-based networks, 106th AES Convention, pre-print No. 4873, May 1999.

[6] HOLLIER (M.P.), HAWKSFORD (M.O.), GUARD (D.R.), Characterisation of communications systems using a speech-like test stimulus, Journal of the AES, 41 (12), 1008-1021, December 1993.

[7] ZWICKER (Feldtkeller): *Das Ohr als Nachrichtenempfänger, S. Hirzel Verlag,* Stuttgart, 1967.

[8] Draft ITU-T recommendation P.862, "Telephone transmission quality, telephone installations, local line networks - Methods for objective and subjective assessment of quality - Perceptual evaluation of speech quality (PESQ), an objective method for en-to-end speech qualtity assessment of narrow-bank telephone networks and speech codecs", ITU-T 02.2001

[9] European patent application EP01200945, Koninklijke KPN n.v.

Claims

1. Method for measuring the transmission quality of an audio system, an input signal (X) being entered into the audio system, resulting in an output signal (Y), output by the audio system, the input signal and the output signal being processed preferably compared, comprising that the output signal and/or the input signal of the audio system are scaled in a way that small deviations of the power are compensated, while larger deviations are compensated partially, dependent on the power ratio.

2. Method according to claim 1, comprising that
a compensation ratio F is calculated from the power representations PX and PY resp. of said input signal (X) and output signal (Y), F being equal to the ratio PX/PY;
a clipped ratio C being calculated, C being equal to a first clipping value mm for F<mm, or C being equal to a second clipping value MM for F>MM, or otherwise C being equal to F;
a softscale ratio S being calculated from a first scaling factor m and a second scaling factor M, with $mm < m \leq 1$ and $MM > M \geq 1$, S being equal to $C^{a} + C - C(m)^{a-1}$ for C < m, 'a' being a first tuning parameter being set to a value

> 0 and < 1, or S being equal to $C^a + C - C(M)^{a-1}$ for C > M, or otherwise S being equal to C.

3. Method according to claim 1, comprising that an artificial reference speech signal is created, for which the noise levels as present in the original input speech signal are lowered by a scaling factor that depends on the local level of the noise in this input.

4. Method according to claim 3, comprising that the difference D(f)n in internal representations LX(f)n and LY(f)n resp. of said input signal (X) and output signal (Y) in the time-frequency plane is set to be equal to $|LY(f)n - LX(f)n^b/K^{b-1}|$ for LX(f)n < K, or to be equal to |LY(f)n - LX(f)n| for LX(f)n ≥ K, b being a second tuning parameter being set to a value > 1 and K being a low level noise power criterion value, representing a desired low level noise power criterion.

5. Method according to claim 3, comprising that a difference D(f)n in internal representations LX(f)n and LY(f)n resp. of said input signal (X) and output signal (Y) in the time-frequency plane is set to be equal to $|LY(f)n - LX(f)n^b/K^{b-1}|$ for LX(t) < K' or be equal to |LY(f)n - LX(f)n| for LX(t) ≥ K', b being a second tuning parameter being set to a value > 1 and K' being a low level noise power criterion value, representing the desired low level noise power criterion.

6. System for measuring the transmission quality of an audio system, an input signal (X) being entered into the audio system, resulting in an output signal (Y), output by the audio system, the input signal and the output signal being mutually processed preferably compared, comprising scaling means for scaling the output signal and/or the input signal of the audio system in a way that small deviations of the power are compensated, while larger deviations are compensated partially, dependent on the power ratio.

7. System according to claim 6, comprising means for calculating a compensation ratio F from the power representations PX and PY resp. of said input signal (X) and output signal (Y), F being equal to the ratio PX/PY; comprising means for calculating a clipped ratio C, being equal to a first clipping value mm for F<mm, or C being equal to a second clipping value MM for F>MM, or otherwise C being equal to F; and comprising means for calculating a softscale ratio S from a first scaling factor m and a second scaling factor M, with mm < m ≤ 1 and MM > M ≥ 1, S being equal to $C^a + C - C(m)^{a-1}$ for C < m, 'a' being a first tuning parameter being set to a value > 0 and < 1, or S being equal to $C^a + C - C(M)^{a-1}$ for C > M, or otherwise S being equal to C.

8. System according to claim 6, comprising means for the creation of an artificial reference speech signal, for which the noise levels as present in the original input speech signal are lowered by a scaling factor that depends on the local level of the noise in this input.

9. System according to claim 8, comprising means for setting the difference D(f)n in internal representations LX(f)n and LY(f)n resp. of said input signal (X) and output signal (Y) in the time-frequency plane to be equal to $|LY(f)n - LX(f)n^b/K^{b-1}|$ for LX(f)n < K, or to be equal to |LY(f)n - LX(f)n| for LX(f)n ≥ K, b being a second tuning parameter being set to a value > 1 and K being a low level noise power criterion value, representing a desired low level noise power criterion.

10. System according to claim 8, comprising means for setting a difference D(f)n in internal representations LX(f)n and LY(f)n resp. of said input signal (X) and output signal (Y) in the time-frequency plane to be equal to $|LY(f)n - LX(f)n^b/K^{b-1}|$ for LX(t) < K' or be equal to |LY(f)n - LX(f)n| for LX(t) ≥ K', b being a second tuning parameter being set to a value > 1 and K' being a low level noise power criterion value, representing the desired low level noise power criterion.

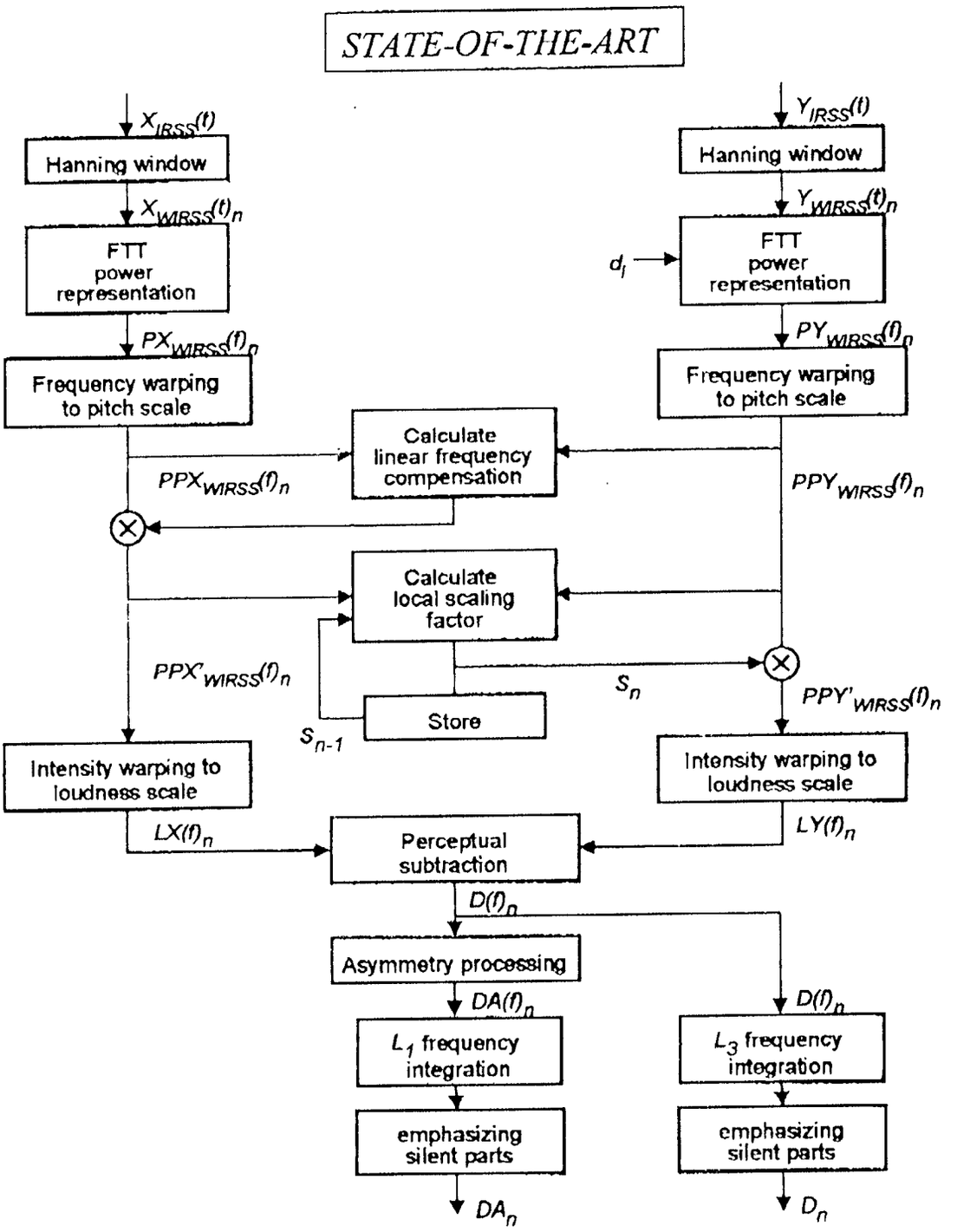
STATE-OF-THE-ART
$X_{IRSS}(t)$
Hanning window
$X_{WIRSS}(t)_n$
FTT power representation
$PX_{WIRSS}(f)_n$
Frequency warping to pitch scale
$Y_{IRSS}(t)$
Hanning window
$Y_{WIRSS}(t)_n$
d_f
FTT power representation
$PY_{WIRSS}(f)_n$
Frequency warping to pitch scale
Calculate linear frequency compensation
$PPX_{WIRSS}(f)_n$
$PPY_{WIRSS}(f)_n$
Calculate local scaling factor
$PPX'_{WIRSS}(f)_n$
S_n
$PPY'_{WIRSS}(f)_n$
Store
S_{n-1}
Intensity warping to loudness scale
Intensity warping to loudness scale
$LX(f)_n$
$LY(f)_n$
Perceptual subtraction
$D(f)_n$
Asymmetry processing
$DA(f)_n$
$D(f)_n$
L_1 frequency integration
L_3 frequency integration
emphasizing silent parts
emphasizing silent parts
DA_n
D_n

FIG. 1

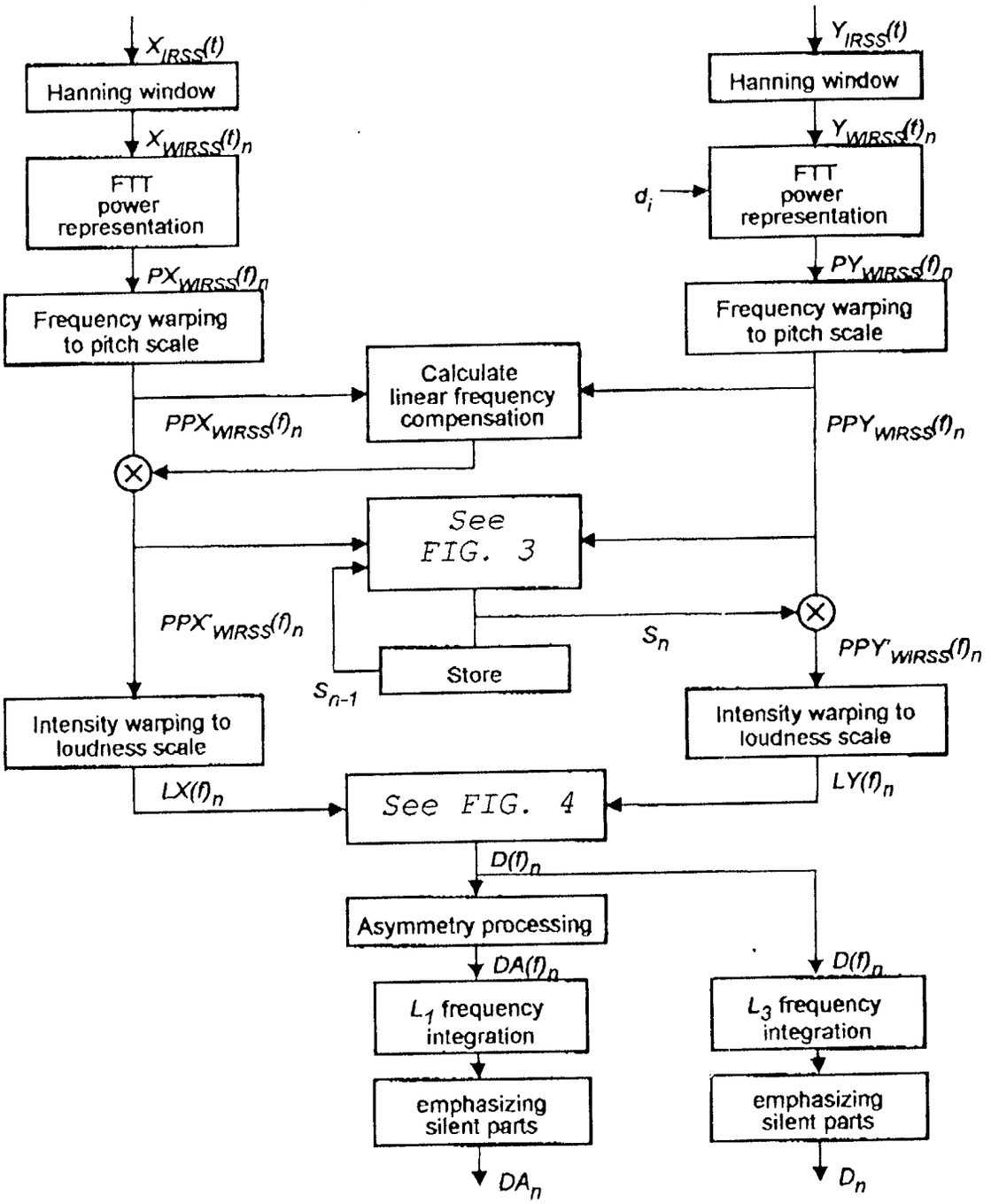
$X_{IRSS}(t)$
Hanning window
$X_{WIRSS}(t)_n$
FTT power representation
$PX_{WIRSS}(f)_n$
Frequency warping to pitch scale
$Y_{IRSS}(t)$
Hanning window
$Y_{WIRSS}(t)_n$
d_i
FTT power representation
$PY_{WIRSS}(f)_n$
Frequency warping to pitch scale
Calculate linear frequency compensation
$PPX_{WIRSS}(f)_n$
$PPY_{WIRSS}(f)_n$
See FIG. 3
$PPX'_{WIRSS}(f)_n$
S_n
$PPY'_{WIRSS}(f)_n$
Store
S_{n-1}
Intensity warping to loudness scale
Intensity warping to loudness scale
$LX(f)_n$
$LY(f)_n$
See FIG. 4
$D(f)_n$
Asymmetry processing
$DA(f)_n$
$D(f)_n$
L_1 frequency integration
L_3 frequency integration
emphasizing silent parts
emphasizing silent parts
DA_n
D_n

FIG. 2

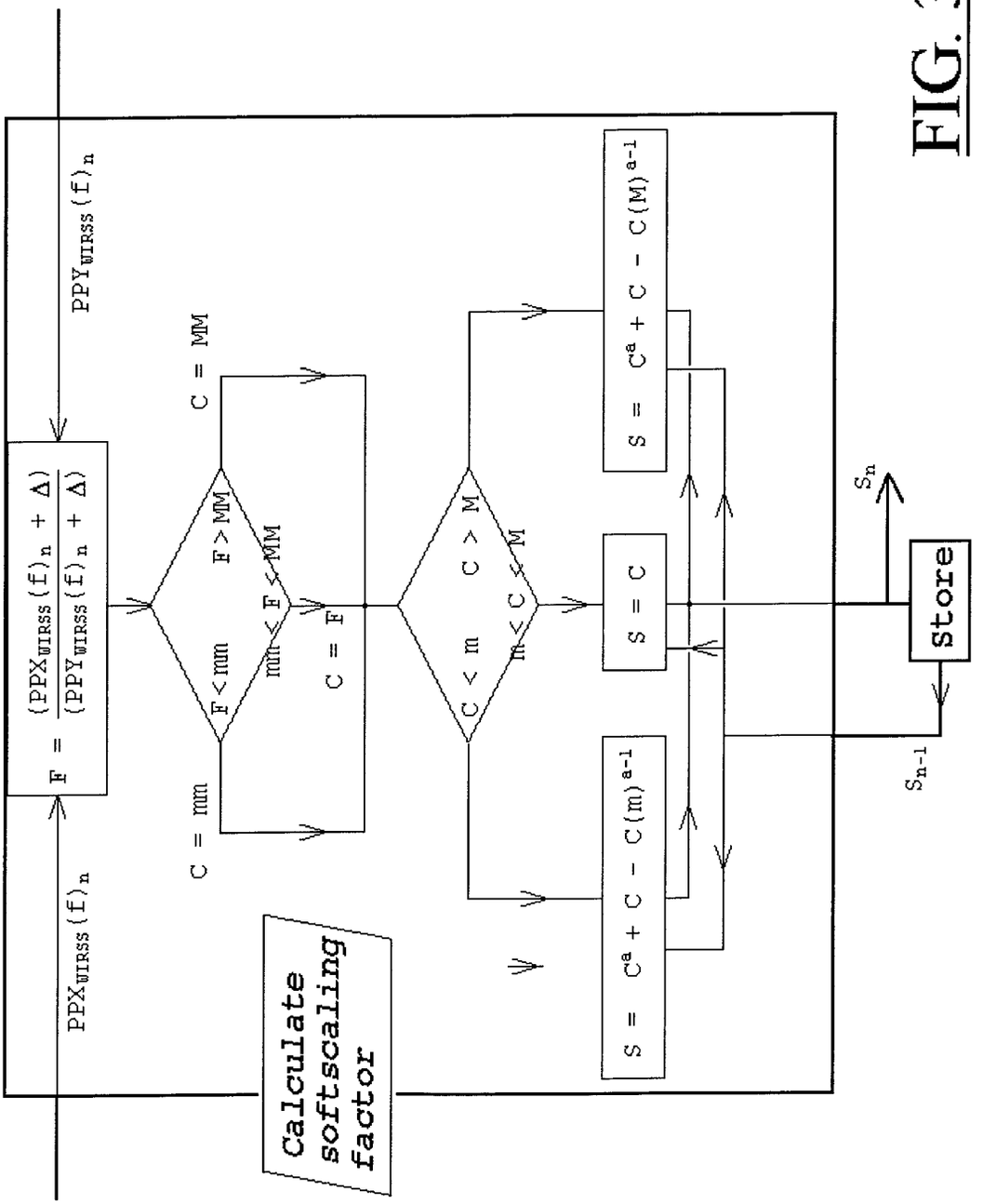
PPY WIRSS (f) n
PPX WIRSS (f) n
F = (PPX WIRSS (f) n + Δ) / (PPY WIRSS (f) n + Δ)
F < mm
F > MM
mm < F < MM
C = mm
C = MM
C = F
C < m
C > M
m < C < M
S = C
S = C^a + C - C(m)^a-1
S = C^a + C - C(M)^a-1
Store
S n
S n-1
Calculate softscaling factor

FIG. 3

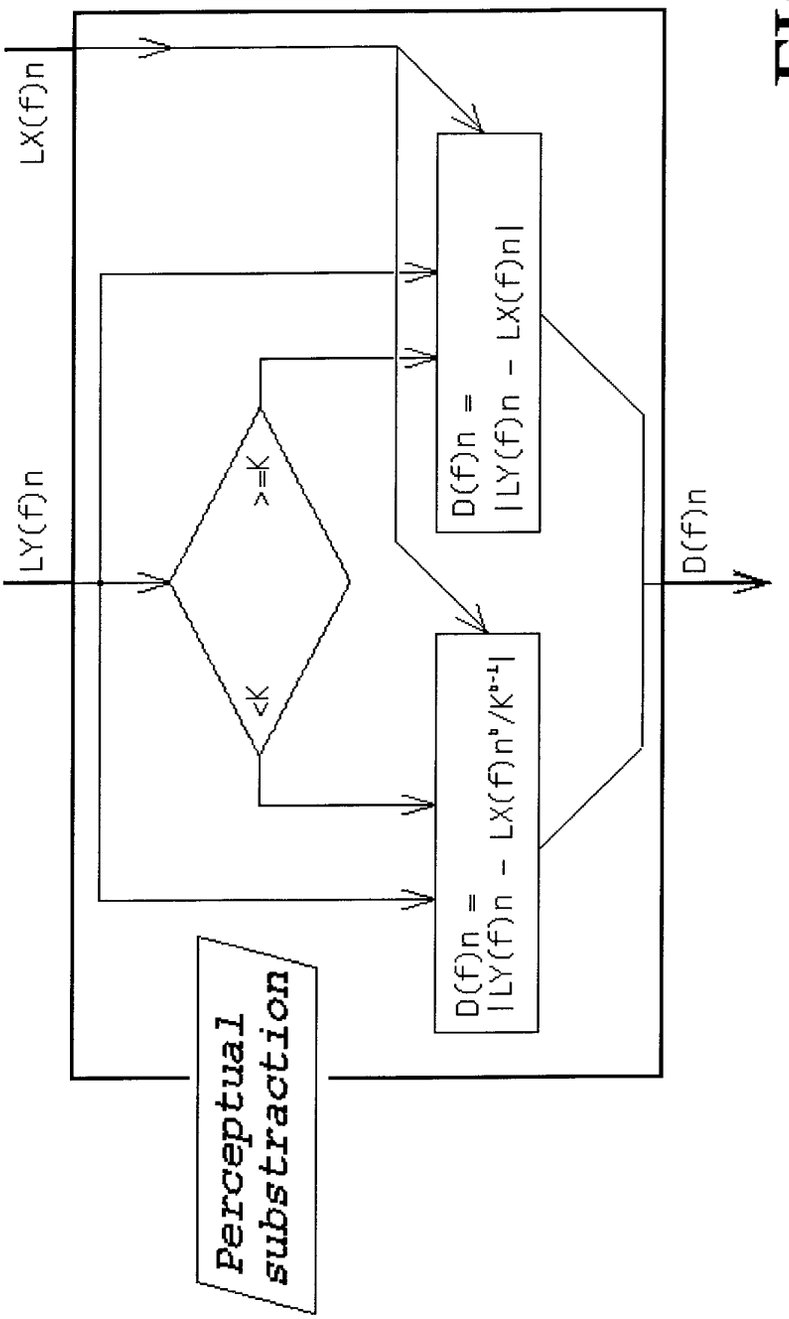
LX(f)n
LY(f)n
>=K
<K
D(f)n =
|LY(f)n - LX(f)n|
D(f)n =
|LY(f)n - LX(f)n^b/K^(b-1)|
D(f)n
Perceptual substraction

FIG. 4

European Patent Office

EUROPEAN SEARCH REPORT

Application Number
EP 02 07 5973

DOCUMENTS CONSIDERED TO BE RELEVANT

Category	Citation of document with indication, where appropriate, of relevant passages	Relevant to claim	CLASSIFICATION OF THE APPLICATION (Int.Cl.7)
A	BEERENDS J.G., HEKSTRA A.P., RIX A.W. AND HOLLIER M.P.: "Perceptual Evaluation of Speech Quality (PESQ), the new ITU standard for end-to-end speech quality assessment. Part II - Psychoacoustic model" WWW.PSYTECHNICS.COM/PAPERS/, June 2001 (2001-06), pages 1-27, XP002206026 * section "3.1 Calibration" * * section "3.7 Compensation of the Time Varying Gain" * ---	1-10	G10L19/00
A	RIX A W ET AL: "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs" 2001 IEEE INTERNATIONAL CONFERENCE ON ACOUSTICS, SPEECH, AND SIGNAL PROCESSING. PROCEEDINGS (CAT. NO.01CH37221), vol. 2, 7 - 11 May 2001, pages 749-752 , XP002187839 SALT LAKE CITY, UT, USA, Piscataway, NJ, USA, IEEE, USA ISBN: 0-7803-7041-4 * the whole document * ---	1-10	TECHNICAL FIELDS SEARCHED (Int.Cl.7) G10L
A	JOHN ANDERSON: "Methods for Measuring Perceptual Speech Quality passage" METHODS FOR MEASURING PERCEPTUAL SPEECH QUALITY, XX, XX, 1 March 2001 (2001-03-01), pages 1-34, XP002172414 * page 25 - page 29 * --- -/--	1-10	

The present search report has been drawn up for all claims

Place of search	Date of completion of the search	Examiner
THE HAGUE	15 July 2002	Quélavoine, R

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

EPO FORM 1503 03.82 (P04C01)

European Patent Office

EUROPEAN SEARCH REPORT

Application Number
EP 02 07 5973

DOCUMENTS CONSIDERED TO BE RELEVANT

Category	Citation of document with indication, where appropriate, of relevant passages	Relevant to claim	CLASSIFICATION OF THE APPLICATION (Int.Cl.7)
A	BEERENDS J.G., HEKSTRA A.P., RIX A.W. AND HOLLIER M.P.: "Perceptual Evaluation of Speech Quality (PESQ), the new ITU standard for end-to-end speech quality assessment. Part I - Time alignment" WWW.PSYTECHNICS.COM/PAPERS/, June 2001 (2001-06), pages 1-9, XP002206027 * abstract * ---	1,6	

TECHNICAL FIELDS SEARCHED (Int.Cl.7)

The present search report has been drawn up for all claims

Place of search	Date of completion of the search	Examiner
THE HAGUE	15 July 2002	Quélavoine, R

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

EPO FORM 1503 03.82 (P04C01)