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

(11) **EP 2 048 657 B1**

(12) **EUROPEAN PATENT SPECIFICATION**

(45) Date of publication and mention
of the grant of the patent:
09.06.2010 Bulletin 2010/23

(51) Int Cl.:
G10L 19/00 *(2006.01)*

(21) Application number: **07019894.0**

(22) Date of filing: **11.10.2007**

(54) **Method and system for speech intelligibility measurement of an audio transmission system**

Verfahren und System zur Messung der Sprachverständlichkeit eines Tonübertragungssystems

Procédé et système de mesure de l'intelligibilité de la parole d'un système de transmission audio

(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 MT NL PL PT RO SE SI SK TR

(43) Date of publication of application:
15.04.2009 Bulletin 2009/16

(73) Proprietors:
- **Koninklijke KPN N.V.**
  2516 CK The Hague (NL)
- **Nederlandse Organisatie voor Toegepast-Natuurwetenschappelijk Onderzoek TNO**
  2628 VK Delft (NL)

(72) Inventors:
- **Beerends, John Gerard**
  4585 PG Hengstdijk (NL)
- **Van Buuren, Ronald Alexander**
  2411 JA Bodegraven (NL)
- **Van Vugt, Jeroen Martijn**
  2518 VD The Hague (NL)

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

(56) References cited:
EP-A- 0 809 236 EP-A- 1 241 663

- **BEERENDS J G ET AL: "PERCEPTUAL EVALUATION OF SPEECH QUALITY (PESQ) THE NEW ITU STANDARD FOR END-TO-END SPEECH QUALITY ASSESSMENT PART II-PSYCHOACOUSTIC MODEL" JOURNAL OF THE AUDIO ENGINEERING SOCIETY, AUDIO ENGINEERING SOCIETY, NEW YORK, NY, US, vol. 50, no. 10, October 2002 (2002-10), pages 765-778, XP001245918 ISSN: 1549-4950**
- **APPEL R ET AL: "ON THE QUALITY OF HEARING ONE'S OWN VOICE" JOURNAL OF THE AUDIO ENGINEERING SOCIETY, AUDIO ENGINEERING SOCIETY, NEW YORK, NY, US, vol. 50, no. 4, April 2002 (2002-04), pages 237-248, XP001130079 ISSN: 1549-4950**

Printed by Jouve, 75001 PARIS (FR)

Description

Field of the invention

[0001] The present invention relates to a method for measuring the speech intelligibility of an audio transmission system, an input signal *X(t)* being entered into the system, resulting in an output signal Y(t), in which both the input signal *X(t)* and the output signal Y(t) are processed. In a further aspect, the present invention relates to a processing system for measuring the intelligibility of a degraded output signal Y(t) from an audio transmission system in response to a reference input signal *X(t).*

Prior art

[0002] A related method and system are known from ITU-T recommendation P.862, "Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs", ITU-T 02.2001 [3].

[0003] Also, the article by J. Beerends et al. "Perceptual Evaluation of Speech Quality (PESQ), the new ITU standard for end-to-end speech quality assessment, Part II - Psychoacoustic model," J. Audio Eng. Soc., vol. 50, pp. 765-778 (2002 Oct.), describes such a method and system [2].

[0004] The present invention is a further development of the idea that speech and audio intelligibility measurement should be carried out in the perceptual domain. In general this idea results in a system that compares a reference speech signal with a distorted signal that has passed through the system under test. By comparing the internal perceptual representations of these signals, estimation can be made about the perceived intelligibility. The latest technology relating to similar quality measurement in this field can be found in references [1] ...[11]. All currently available systems suffer from the fact that speech intelligibility cannot be measured. In a database that was constructed with a CVC (Consonant Vowel Consonant) identification task, the correlation between CVC correct scores and raw PESQ scores was below 0.6. The currently best method for measuring speech intelligibility is the STI (Speech Transmission Index), see references [12] ...[15]. However the STI method uses a modulated noise, speech like, test signal and can only be used under a limited set of distortions.

Summary of the invention

[0005] The present invention seeks to provide a new measurement method and apparatus for measuring the intelligibility of speech as output in a speech/audio communication system.

[0006] According to the present invention, a method according to the preamble defined above is provided, in which the method comprises:

- preprocessing of the input signal (*X(t)*) and output signal (*Y(t)*) to obtain pitch power densities ($PPX(f)_n$, $PPY(f)_n$) for the respective signals, comprising pitch power density values for cells in the frequency (*f*) and time *(n)* domain;
- compensating the pitch power densities to obtain compensated pitch power densities ($PPX'(f)_n$, $PPY'(f)_n$);
- transforming the compensated pitch power densities ($PPX'(f)_n$, $PPY'(f)_n$) in loudness densities ($LX(f)_n$, $LY(f)_n$);
- perceptual subtraction of the loudness densities ($LX(f)_n$, $LY(f)_n$) to obtain a disturbance density function ($D(f)_n$);
- correction of the disturbance density function ($D(f)_n$) by multiplying the disturbance density function ($D(f)_n$) with a correction function for each frame derived from a correlation calculation of the compensated pitch power density ($PPX'(f)_n$) associated with the input signal *(X(t))* of a present frame *(n)* and an independent previous frame to obtain a corrected disturbance density function ($D'(f)_n$); and
- aggregating the corrected disturbance density function ($D'(f)_n$) over frequency and time to obtain a measure (I) for the speech intelligibility of the output signal (*Y(t)*).

[0007] With the term independent previous frame it is meant to have a previous frame which does not have any overlap with the present frame. E.g. the frames may have a 50% overlap, in which case the compensated pitch power density associated with the present frame n is correlated with compensated pitch power density associated with the second previous frame *n*-2.

[0008] By correcting the disturbance density function in the described manner, the correlation between the measure for the speech intelligibility as calculated by the present method embodiment and actual speech intelligibility scores are improved. The present invention is based on the insight that when two frames in a speech signal are alike, degradations as found by the prior art PESQ method are causing less decrease in intelligibility than predicted. When a subject is hearing a sound a second time, the subject is able to better understand it than the first time the (same) sound is heard.

[0009] In a further embodiment, the correction function (frameCorTimeOrg(n)) is calculated according to:

$$\text{frameCorTimeOrg(n)} =$$

$$\text{frameCorTimeOrg(n)} = \text{FrequencybandCorrelation}(PPX'(f)_n, PPX'(f)_{n-2})$$

In the existing PESQ method, such a feature allows to easy amend the method to the changed insight for predicting speech intelligibility scores.

[0010] In an even further embodiment, correlation calculation is executed over a frequency domain range from a low frequency limit to a high frequency limit, such as the range from 100...3500Hz. As this corresponds to the general speech frequency range, it is sufficient to restrict the calculations to this range for predicting intelligibility of a sound signal.

[0011] The correction function may be limited to a value less or equal to 1.0, according to the rules:

```
if frameCorTimeOrg(n) < 0.0
  frameCorrelationTimeCompensation = 1.0
else
  frameCorrelationTimeCompensation = 1.0 - ( frameCorTimeOrg(n) )^k,
k being a predetermined power value.
```

[0012] The predetermined power value may be larger than 1.0, e.g. between 10 and 20. In this manner, the method incorporates that for low correlations, the impact on the intelligibility score is marginal, and only correlations close to 1.0 are included more pronounced as their impact is significant.

[0013] In an even further embodiment, the correction function is limited to a value larger than or equal to a lower limit value, e.g. 0.4. This assures that the corrections as applied to the disturbance density functions are not influenced too heavily for strong correlating frames.

[0014] As in the prior art PESQ method, the (corrected) disturbance density function is aggregated over the frequency and time domain, to yield a measure in the form of a value. From this measure, the speech intelligibility may be provided with a score, e.g. using a mapping similar to a CVC intelligibility score.

[0015] Specific for the measurement of intelligibility, the aggregation functions over frequency and time are adapted. In a further embodiment, the corrected disturbance density function $D'(f)_n$ is aggregated over frequency using a low norm factor (Lq), in which the low norm factor (Lq) has a value of less than or equal to 2, and aggregated over time using a high norm factor (L_p), in which the high norm factor (L_p) has a value of greater than or equal to 6.

[0016] In a further embodiment, the method further comprises calculating a difference between two intelligibility score measures (I), in which the intelligibility score measures (I) are calculated using different norm factors, the norm factors being less than or equal to 3. This provides an even further improved intelligibility score measurement, which is even closer to actual subjective tests.

[0017] In a further aspect, the present invention relates to a processing system as described above, comprising a processor connected to the audio transmission system for receiving the reference input signal *X(t)* and the degraded output signal Y(t), in which the processor is arranged for outputting a measure I for the speech intelligibility of the output signal *Y(t),* and for executing the steps of the method according to any one of the present method embodiments.

[0018] In an even further aspect, the present invention relates to a computer program product comprising computer executable software code, which when loaded on a processing system, allows the processing system to execute the method according to any one of the present method embodiments.

Short description of drawings

[0019] The present invention will be discussed in more detail below, using a number of exemplary embodiments, with reference to the attached drawings, in which

Fig. 1 shows a block diagram of an application of the present invention;
Fig. 2 shows a flows chart of the implementation of an embodiment of the present invention.

Detailed description of exemplary embodiments

[0020] During the past decades a number of measurement techniques have been developed that allow to quantify the quality of audio devices in a way that closely copies human perception. The advantage of these methods over classical methods that quantify the quality in terms of system parameters like frequency response, noise, distortion, etc is the high correlation between subjective measurements and objective measurements. With this perceptual approach a series of audio signals is input into the system under test and the degraded output signal is compared with the original input to the system on the basis of a model of human perception. On the basis of a set of comparisons the intelligibility of the system under test can be quantified.

[0021] The perceptual model uses the basic features of the human auditory system to map both the original input and the degraded output onto an internal representation. If the difference in this internal representation is zero the system under test is transparent for the human observer representing a perfect system under test (from the perspective of perceived audio intelligibility). If the difference is larger then zero it is mapped to an intelligibility number using a cognitive model, allowing quantifying the perceived degradation in the degraded output signal.

[0022] Fig. 1 shows schematically a known set-up of an application of an objective measurement technique which is based on a model of human auditory perception and cognition, and which follows the ITU-T Recommendation P.862 (see reference [3]), for estimating the perceptual quality of speech links or codecs, which can also be applied for the present invention relating to intelligibility measurement. The acronym used for this technique or device is PESQ (Perceptual Evaluation of Speech Quality). It comprises a system or telecommunications network under test 10, hereinafter referred to as system 10, and a measurement device 11 for the perceptual analysis of speech signals offered. A speech signal $X_0(t)$ is used, on the one hand, as an input signal of the system 10 and, on the other hand, as a first input signal X(t) of the device 11. An output signal Y(t) of the system 10, which in fact is the speech signal $X_0(t)$ affected or degraded by the system 10, is used as a second input signal of the measurement device 11. An output signal I of the measurement device 11 represents an estimate of the perceptual intelligibility of the speech link through the system 10.

[0023] The measurement device 11 may be implemented as a processing system comprising a dedicated signal processing unit, e.g. having one or more (digital) signal processors, or a general purpose processing system having one or more processors under the control of a software program comprising computer executable code. The device 11 is provided with suitable input and output modules and further supporting elements for the processors, such as memory, as will be clear to the skilled person.

[0024] Since the input end and the output end of a speech link (shown as the system 10 in Fig. 1), particularly in the event it runs through a telecommunications network, are remote, use is made in most cases of speech signals X(t) stored on data bases for the input signals of the measurement device 11. Here, as is customary, speech signal is understood to mean each sound basically perceptible to the human hearing, such as speech and tones. The system under test 10 may of course also be a simulation system, which e.g. simulates a telecommunications network.

[0025] The present invention solves the problem of low correlation between the PESQ scores and speech intelligibility scores by an additional new processing step for calculating the internal representation of the speech signal. It uses PESQ P.862.1 (reference [4]) and P.862.2 (reference [5]) as the starting point for an algorithm that can predict the perceived speech intelligibility of a speech fragment. The documents reference [3], [4], and [5] show the general steps of the PESQ method.

[0026] The present method can be used on normal speech material as well as on a short CVC test signal (Consonant Vowel Consonant). This test signal $X_0(t)$ contains a set of short speech fragments, concatenated CVC words as used in speech intelligibility testing, that contains all relevant vowels and consonants, including the relevant transitions, and is put into the system under test 10.

[0027] In Fig. 2, a flow chart is shown in schematic form of an embodiment of the present invention, which may be implemented in the measurement device 11 shown in Fig. 1. The starting processing blocks 21-34, as well as the final blocks 35-37 are the general processing steps applied in PESQ, see reference [3], although it should be noted that other embodiments comprising one or more additional or amended processing steps are possible, to obtain more specialized measuring methods or measuring methods with other objectives. These starting blocks 21-34 will be discussed in short, after which the further processing steps 50-55 of the present method embodiment are discussed in more detail, as well as the final blocks 35-37.

[0028] The first step in the PESQ algorithm is to compensate for the overall gain of the system under test, which is executed in the level and level/time alignment blocks 21, 22. These steps 21, 22 are combined with a global scaling of the signals to a correct overall level in block 27. Both the original X(t) (reference input signal) and degraded (output) signal $Y(t)$ are scaled to the same, constant power level, resulting in signals $X_s(t)$ and $Y_s(t)$.

[0029] Then, these signals are subjected to a windowed fast Fourier transform operation, in respective blocks 23, 24, resulting in the power representation arrays $PX(f)_n$ and $PY(f)_n$. The human ear performs a time-frequency transformation. In PESQ this is modelled by a short term FFT with a Hann window over 32 ms frames. The overlap between successive frames is 50%. 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 frame is discarded in PESQ and all calculations are based on only the power representations $PX(f)_n$ and $PY(f)_n$.

[0030] In the next processing blocks, both power representation arrays $PX(f)_n$ and $PY(f)_n$, are subjected to a frequency warping operation to a pitch scale in processing blocks 25 and 26, respectively. 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 approximates the values given in the literature. The resulting signals are known as the pitch power densities $PPX(f)_n$ and $PPY(f)_n$.

[0031] To deal with the subjective impact of linear distortions as formed in the system under test, a (partial) frequency response compensation is executed in processing block 28. The pitch power densities $PPX(f)_n$ and $PPY(f)_n$ 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 30 dB above 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 20dB. The original pitch power density $PPX(f)_n$ of each frame n is then multiplied with this partial compensation factor to equalise the original to the degraded signal. This results in a filtered version of the original pitch power density $PPX'(f)_n$. This partial compensation is used because severe filtering is disturbing to the listener while mild filtering effects hardly influence the perceived overall quality and intelligibility, especially if no reference is available to the subject. The compensation is carried out on the original signal because the degraded signal is the one that is judged by the subjects in an Absolute Category Rating (ACR) experiment.

[0032] Short-term gain variations are partially compensated by processing the pitch power densities frame by frame, as indicated in processing block 29. For the original and the degraded pitch power densities $(PPX(f)_n$ and $PPY(f)_n$ in the embodiment shown in Fig. 2), 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 time constant of this filter is approximately 16ms. 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 $PPY'(f)_n$.

[0033] After partial compensation for filtering and short-term gain variations in processing blocks 28, the original pitch power densities are transformed to a Sone loudness scale using Zwicker's law in processing block 31.

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

with $P_0(f)$ the absolute hearing threshold and S_1 the loudness scaling factor. In a similar manner, the output (or degraded) pitch power densities $PPY'(f)_n$ is transformed in processing block 32. The resulting two dimensional arrays $LX(f)_n$ and $LY(f)_n$ are called loudness densities.

[0034] The signed difference between the distorted and original loudness density $LX(f)_n$ and $LY(f)_n$ is computed in processing block 34, labelled as perceptual subtraction. 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.

[0035] Masking is modelled by applying a dead zone in each time-frequency cell, as follows. The per cell 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. Next 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.

[0036] 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 function as a function of time (frame number n) and frequency, $D(f)_n$.

[0037] According to the present invention embodiments an additional processing step is introduced to obtain a better correlation between speech intelligibility scores and the final PESQ score I. The present invention embodiments use PESQ P.862.1 and P.862.2 (see reference [4] and [5]) as the starting point for an algorithm that can predict the perceived speech intelligibility of a speech fragment. The method can be used on normal speech material as well as on a short CVC test signal (Consonant Vowel Consonant). This test signal contains a set of short speech fragments, concatenated CVC words as used in speech intelligibility testing, that contains all relevant vowels and consonants, including the relevant transitions, and is put into the system under test.

[0038] The additional processing, which is shown schematically in Fig. 2 as processing blocks 50-55, is based on the insight that when two frames (frame length about 30 ms) within a speech signal are alike, i.e. a high correlation between their pitch power density functions, then the degradations as found by PESQ in the second frame are causing less decrease in intelligibility then predicted on the basis of the PESQ disturbance. When a sound is repeated subjects are able to better understand its meaning then when they hear the sound for the first time.

[0039] To quantify this effect the symmetric disturbance function $D(f)_n$ as defined in PESQ is compensated for each time frame n with a correction function (frameCorrelationTimeCompensation) that is derived from the correlation between the current time frame pitch power density $PPX'(f)_n$, and the previous independent time frame pitch power density $PPX'(f)_{n-2}$ of the reference input file.

[0040] With the term independent previous frame it is meant to have a previous frame which does not have any overlap with the present frame. E.g. the frames may be based on 50% overlapped cos^2 windows with index n, in which case the compensated pitch power density associated with the present frame *n* is correlated with compensated pitch power density associated with the second previous frame *n-2*.

[0041] This is calculated according to:

$$\text{frameCorTimeOrg(n)} = \text{FrequencybandCorrelation}(PPX'(f)_n, PPX'(f)_{n-2})$$

In an embodiment, this function is calculated with the frequency index f: e.g. 100 Hz < f < 3500 Hz, as only speech energy is important in the calculation. The present and previous time frame pitch power densities $PPX'(f)_n$, $PPX'(f)_{n-2}$ are stored in associated blocks 51, 52. The correlation calculation is implemented in processing block 50. Then, in processing block 53, the correction function is calculated according to:

```
if frameCorTimeOrg(n) < 0.0
  frameCorrelationTimeCompensation = 1.0
else
  frameCorrelationTimeCompensation = 1.0 - ( frameCorTimeOrg(n) )$^k$;
if frameCorrelationTimeCompensation  < 0.4
      frameCorrelationTimeCompensation = 0.4
```

[0042] The value of the correction function frameCorrelationTimeCompensation is thus limited between a lower limit (in the example shown 0.4) and an upper limit (i.e. 1).

[0043] The predetermined power value k quantifies the point where the frameCorrelationTimeCompensation starts to have an impact. For low correlations the impact is marginal, only when the correlation is close to 1.0 the impact is significant. This leads to an optimal k>>1.0. In a specifically advantageous embodiment, the value k lies between 10 and 20.

[0044] In an embodiment of the present invention, first a speech signal *X(t)* containing the speech fragments with which the system under test 10 has to be evaluated is inputted to the measurement system 11. Next the internal representation as described in PESQ P.862 [3], [4], [5] is calculated by the measurement system 11 for both the reference input *X(t)* and the degraded output *Y(t)* and from that the symmetric disturbance density $D(f)_n$ (see above) and an asymmetric disturbance density $DA(f)_n$ (see reference [3]). In the current best practice only the symmetric disturbance $D(f)_n$ is used in combination with the frameCorrelationTimeCompensation as described above. For each frame *n* the corrected disturbance density $D'(f)_n$ is calculated from the product of the disturbance density $D(f)_n$ and the frameCorrelationTimeCompensation.

[0045] This corrected disturbance density is then integrated over the frequency, the speech spurts and the complete file length similarly as carried out in PESQ P.862 but with a low norm factor (power factor Lq) over frequency and spurt (e.g. Lq<2, e.g. L_q=1) and a high norm factor (power factor L_p) over time (e.g. L_p>6, e.g. L_p=8).

[0046] In processing block 35, an aggregation of disturbance densities over frequency is performed using the low norm factor Lq according to:

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

with M_n a multiplication factor equal to *((power of original frame + 10^5)/ 10^7)$^{-0.04}$,* 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 are called frame disturbances.

[0047] In processing block 36, an aggregation of the frame disturbances over time is executed similarly using the low norm factor Lq for the speech spurts, and the high norm factor Lp for the aggregation over the entire speech sample.

[0048] In general, the existing PESQ methods also use a time weighting procedure, to account for the fact that disturbances that occur during speech active periods are more disturbing than those that occur during silent intervals:

$$L_p = \left(\frac{1}{N} \sum_{n=1}^{N} disturbance[n]^p \right)^{1/p} ,$$

with N = total number of frames and p>1.0.

Such an Lp weighting emphasizes loud disturbances when compared to a normal, L_1 time averaging, leading to a better correlation between objective and subjective scores The aggregation of frame disturbances over time is carried in a hierarchy of two layers.

[0049] The present invention embodiments are somewhat different from the standard PESQ method (reference [3]). First, the aggregation over frequency is executed using a norm factor equal to 3 instead of the low norm value of 2 in the present embodiment. Furthermore, in the standard PESQ method, the frame disturbance values are aggregated over split second intervals of 20 frames (accounting for the overlap of frames: approx. 320 ms using a norm factor equal to 8. These intervals also overlap 50 per cent and no window function is used. The split second disturbance values are aggregated over the active interval of the speech files (the corresponding frames) now using a norm factor equal to 2.

[0050] As a result, a disturbance indicator *D* is obtained, which can be further mapped onto a final CVC intelligibility score in processing block 37 (the quantity I in Fig. 1).

[0051] The present invention embodiments result in a quantity I that shows a strong correlation with the speech intelligibility of the output speech signal *Y(t).*

[0052] A further improvement can be obtained using an even further embodiment, from calculating the difference between two frequency, spurt, time integrations, both with a low L_p power (<3). In the above example, the integration over frequency, spurt, time integration has been done using 1, 1, and 8 as respective norm factors L_p, L_p, L_q. In this further example, two calculations are made which are then subtracted from each other. E.g., a first calculation is made using 2, 3, 2 as respective norm factors for the integration over frequency, spurt and entire speech sample, and a second calculation using 1, 3, 3, as respective norm factors.

[0053] The present invention has been described above by means of exemplary embodiments. As will be clear to the skilled person, further modifications and alternatives may be used that are within the scope of the appended claims.

References

[0054]

[1] A. W. Rix, M. P. Hollier, A. P. Hekstra and J. G. Beerends, "PESQ, the new ITU standard for objective measurement of perceived speech quality, Part 1 - Time alignment," J. Audio Eng. Soc., vol. 50, pp. 755-764 (2002 Oct.).

[2] J. G. Beerends, A. P. Hekstra, A. W. Rix and M. P. Hollier, "Perceptual Evaluation of Speech Quality (PESQ), the new ITU standard for end-to-end speech quality assessment, Part II-Psychoacoustic model," J. Audio Eng. Soc., vol. 50, pp. 765-778 (2002 Oct.) (equivalent to KPN Research publication 00-32228).

[3] ITU-T Rec. P.862, "Perceptual Evaluation Of Speech Quality (PESQ): An Objective Method for End-to-end Speech Quality Assessment of Narrow-band Telephone Networks and Speech Codecs," International Telecommu-

nication Union, Geneva, Switzerland (2001 Feb.).

[4] ITU-T Rec. P.862.1, "Mapping function for transforming P.862 raw result scores to MOS-LQO," Geneva, Switzerland (2003 Nov.).

[5] ITU-T Rec. P.862.2, "Wideband extension to Recommendation P.862 for the assessment of wideband telephone networks and speech codecs," Geneva, Switzerland (2005 Nov.).

[6] A. P. Hekstra, J. G. Beerends, *"Output power decompensation,"* International patent application 402714; PCT EP02/02342; European patent application 01200945.2, March 2001; Koninklijke PTT Nederland N.V.

[7] J. G. Beerends, *"Frequency dependent frequency compensation,"* International patent application 402736; PCT EP02/05556; European patent application 01203699.2, June 2001; Koninklijke PTT Nederland N.V.

[8] J. G. Beerends, *"Method and system for measuring* a *system's transmission quality," Softscaling,* International patent application 402808; PCT EP03/02058; European patent application 02075973.4-2218, April 2002, Koninklijke PTT Nederland N.V.

[9] J. G. Beerends, *"Filter scale loop,"* International patent application 402894; European patent application EP03075949.2, July 2003, Koninklijke PTT Nederland N.V.

[10] T. Goldstein, J. G. Beerends, H. Klaus and C. Schmidmer, "Draft ITU-T Recommendation P.AAM, An objective method for end-to-end speech quality assessment of narrow-band telephone networks including acoustic terminal (s)," White contribution COM 12-64 to ITU-T Study Group 12, September 2003.

[11] J. G. Beerends, *"Linear frequency distortion impact analyzer,"* International patent application; European patent application EP04077601, November 2004, TNO Nederland N.V.

[12] H.J.M. Steeneken and T. Houtgast, "A physical method for measuring speech-transmission quality," J. Acoust. Soc. Am., vol. 67, pp. 318-326 (1980 Jan.).

[13] IEC, Publication 268-16, Sound system equipment, Part 16: The objective rating of speech intelligibility in auditoria by the RASTI method, 1988

[14] ISO, Technical Report 4870, Acoustics- The construction and calibration of speech intelligibility tests, 1991

[15] H.J.M. Steeneken, "On measuring and predicting speech intelligibility," PhD University of Amsterdam (1992).

[16] J. G. Beerends and J. A. Stemerdink, "A Perceptual Audio Quality Measure based on a psychoacoustic sound representation," J. Audio Eng. Soc., vol. 40, pp. 963-978 (1992 Dec.).

Claims

1. Method for measuring the speech intelligibility of an audio transmission system (10), an input signal ($X(t)$) being entered into the system (10), resulting in an output signal ($Y(t)$), in which both the input signal ($X(t)$) and the output signal ($Y(t)$) are processed, comprising:

- preprocessing of the input signal ($X(t)$) and output signal ($Y(t)$) to obtain pitch power densities ($PPX(f)_n$, $PPY(f)_n$) for the respective signals, comprising pitch power density values for cells in the frequency (f) and time (n) domain;
- compensating the pitch power densities to obtain compensated pitch power densities ($PPX'(f)_n$, $PPY'(f)_n$);
- transforming the compensated pitch power densities ($PPX'(f)_n$, $PPY'(f)_n$) in loudness densities ($LX(f)_n$, $LY(f)_n$);
- perceptual subtraction of the loudness densities ($LX(f)_n$, $LY(f)_n$) to obtain a disturbance density function ($D(f)_n$);

characterized by

- correction of the disturbance density function ($D(f)_n$) by multiplying the disturbance density function ($D(f)_n$) with a correction function for each frame derived from a correlation calculation of the compensated pitch power density ($PPX'(f)_n$) associated with the input signal ($X(t)$) of a present frame (n) and an independent previous frame to obtain a corrected disturbance density function ($D'(f)_n$); and

- aggregating the corrected disturbance density function *($D'(f)_n$)* over frequency and time to obtain a measure (I) for the speech intelligibility of the output signal (*Y(t)*).

2. Method according to claim 1, in which the correction function frameCorTimeOrg(n) is calculated according to:

$$\mathbf{frameCorTimeOrg(n) = FrequencybandCorrelation(}\boldsymbol{PPX'(f)_n, PPX'(f)_{n-2}}\mathbf{).}$$

3. Method according to claim 1 or 2, in which the correlation calculation is executed over a frequency domain range from a low frequency limit to a high frequency limit, such as the range from 100...3500Hz.

4. Method according to any one of claims 1-3, in which the correction function is limited to a value less or equal to 1.0, according to the rules:

```
if frameCorTimeOrg(n) < 0.0

    frameCorrelationTimeCompensation = 1.0

else

    frameCorrelationTimeCompensation = 1.0 - $(\text{frameCorTimeOrg(n)})^k$,
```

k being a predetermined power value.

5. Method according to claim 4, in which the predetermined power value is larger than 1.0, e.g. between 10 and 20.

6. Method according to claim 4 or 5, in which the correction function is limited to a value larger than or equal to a lower limit value, e.g. 0.4.

7. Method according to any one of claims 1-6, in which the corrected disturbance density function ($D'(f)_n$) is aggregated over frequency using a low norm factor (L_q), in which the low norm factor (L_q) has a value of less than or equal to 2, and aggregated over time using a high norm factor (L_p), in which the high norm factor (L_p) has a value of greater than or equal to 6.

8. Method according to any one of claims 1-6, in which the method further comprises calculating a difference between two intelligibility score measures (I), in which the intelligibility score measures (I) are calculated using different norm factors, the norm factors being less than or equal to 3.

9. A processing system for measuring the intelligibility of a degraded output signal (*Y(t)*) from an audio transmission system (10) in response to a reference input signal (*X*(*t*)), comprising a measurement device (11) connected to the audio transmission system (10) for receiving the reference input signal (*X(t)*) and the degraded output signal (*Y*(*t*)), in which the measurement device (11) is arranged for outputting a measure (I) for the speech intelligibility of the output signal (*Y*(*t*)), and for executing the steps of the method according to any one of the claims 1-8.

10. Computer program product comprising computer executable software code, which when loaded on a processing system, allows the processing system to execute the method according to any one of the claims 1-8.

Patentansprüche

1. Verfahren zur Messung der Sprachverständlichkeit eines Audioübertragungssystems (10), wobei ein Eingangssignal (*X*(*t*)) in das System (10) eingegeben wird und zu einem Ausgangssignal (*Y*(*t*)) führt, wobei sowohl das Eingangssignal (*X*(*t*)) als auch das Ausgangssignal (*Y*(*t*)) verarbeitet werden, mit den folgenden Schritten:

- Vorverarbeiten des Eingangssignal (*X*(*t*)) und des Ausgangssignal (*Y*(*t*)), um Tonhöhenleistungsdichten ($PPX(f)_n$, $PPY(f)_n$) für die jeweiligen Signale zu erhalten, die Tonhöhenleistungsdichtewerte für Zellen im Frequenzbereich (*f*) und im Zeitbereich (*n*) umfassen;

- Kompensieren der Tonhöhenleistungsdichten, um kompensierte Tonhöhenleistungsdichten ($PPX'(f)_n$, $PPY'(f)_n$) zu erhalten;
- Transformieren der kompensierten Tonhöhenleistungsdichten ($PPX'(f)_n$, $PPY'(f)_n$) zu Lautheitsdichten ($LX(f)_n$, $LY(f)_n$) ;
- wahrnehmungsbezogene Subtraktion der Lautheitsdichten ($LX(f)_n$, $LY(f)_n$), um eine Störungsdichtefunktion ($D(f)_n$);

gekennzeichnet durch

- Korrektur der Störungsdichtefunktion ($D(f)_n$) durch Multiplizieren der Störungsdichtefunktion ($D(f)_n$) mit einer Korrekturfunktion für jeden Rahmen, der aus einer Korrelationsberechnung der kompensierten Tonhöhenleistungsdichte ($PPX'(f)_n$) abgeleitet wird, die mit dem Eingangssignal ($X(t)$) eines derzeitigen Rahmens (n) und eines unabhängigen vorherigen Rahmens assoziiert ist, um eine korrigierte Störungsdichtefunktion ($D'(f)_n$) zu erhalten; und
- Aggregieren der korrigierten Störungsdichtefunktion ($D'(f)_n$) über Frequenz und Zeit, um ein Maß (I) für die Sprachverständlichkeit des Ausgangssignal ($Y(t)$) zu erhalten.

2. Verfahren nach Anspruch 1, wobei die Korrekturfunktion frameCorTimeOrg(n) folgendermaßen berechnet wird:

$$\mathtt{frameCorTimeOrg(n)} = \mathtt{FrequencybandCorrelation}\,(PPX'(f)_n,\ PPX'(f)_{n-2}).$$

3. Verfahren nach Anspruch 1 oder 2, wobei die Korrelationsberechnung über einen Frequenzbereichsumfang von einer unteren Frequenzgrenze zu einer oberen Frequenzgrenze, wie zum Beispiel den Umfang von 100...3500 Hz, ausgeführt wird.

4. Verfahren nach einem der Ansprüche 1-3, wobei die Korrekturfunktion gemäß den folgenden Regeln auf einen Wert kleiner oder gleich 1,0 begrenzt wird:

```
if frameCorTimeOrg(n)<0,0
frameCorrelationTimeCompensation = 1,0
else
frameCorrelationTimeCompensation = 1,0
$(\text{frameCorTimeOrg(n)})^k$,
wobei k ein vorbestimmter Leistungswert ist.
```

5. Verfahren nach Anspruch 4, wobei der vorbestimmte Leistungswert größer als 1,0 ist, z.B. zwischen 10 und 20.

6. Verfahren nach Anspruch 4 oder 5, wobei die Korrekturfunktion auf einen Wert größer oder gleich einem Untergrenzenwert, z.B. 0,4, begrenzt wird.

7. Verfahren nach einem der Ansprüche 1-6, wobei die korrigierte Störungsdichtefunktion ($D'(f)_n$) unter Verwendung eines unteren Normfaktors (L_q) über die Frequenz aggregiert wird, wobei der untere Normfaktor (L_q) einen Wert kleiner oder gleich 2 aufweist, und unter Verwendung eines oberen Normfaktors (L_p) über die Zeit aggregiert wird, wobei der obere Normfaktor (L_p) einen Wert größer oder gleich 6 aufweist.

8. Verfahren nach einem der Ansprüche 1-6, wobei das Verfahren ferner das Berechnen einer Differenz zwischen zwei Verständlichkeitsbewertungsmaßen (I) umfasst, wobei die Verständlichkeitsbewertungsmaße (I) unter Verwendung verschiedener Normfaktoren berechnet werden, wobei die Normfaktoren kleiner oder gleich 3 sind.

9. Verarbeitungssystem zur Messung der Verständlichkeit eines verschlechterten Ausgangssignals ($Y(t)$) aus einem Audioübertragungssystem (10) als Reaktion auf ein Referenzeingangssignal ($X(t)$), umfassend eine mit dem Audioübertragungssystem (10) verbundene Messeinrichtung (11) zum Empfangen des Referenzeingangssignals ($X(t)$) und des verschlechterten Ausgangssignals ($Y(t)$), wobei die Messeinrichtung (11) dafür ausgelegt ist, ein Maß (I) für die Sprachverständlichkeit des Ausgangssignals ($Y(t)$) auszugeben und die Schritte des Verfahrens nach einem der Ansprüche 1-8 auszuführen.

10. Computerprogrammprodukt mit computerausführbarem Softwarecode, der, wenn er auf ein Verarbeitungssystem geladen wird, es dem Verarbeitungssystem erlaubt, das Verfahren nach einem der Ansprüche 1-8 auszuführen.

Revendications

1. Procédé de mesure de l'intelligibilité de la parole d'un système de transmission audio (10), un signal d'entrée *(X(t))* étant entré dans le système (10), produisant un signal de sortie *(Y(t)),* dans lequel le signal d'entrée *(X(t))* et le signal de sortie *(Y(t))* sont tous les deux traités, comprenant :

- le prétraitement du signal d'entrée *(X(t))* et du signal de sortie *(Y(t))* pour obtenir des densités de puissance de hauteur tonale $(PPX(f)_n, PPY(f)_n)$ des signaux respectifs, comprenant des valeurs de densité de puissance de hauteur tonale pour les cellules dans le domaine de fréquence *(f)* et de temps *(n) ;*
- la compensation des densités de puissance de hauteur tonale pour obtenir des densités de puissance de hauteur tonale compensées $(PPX'(f)_n, PPY'(f)_n)$ *;*
- la transformation des densités de puissance de hauteur tonale compensées $(PPX'(f)_n, PPY'(f)_n)$ en densités de niveau sonore $(LX(f)_n, LY(f)_n)$ *;*
- la soustraction perceptive des densités de niveau sonore $(LX(f)_n, LY(f)_n)$ pour obtenir une fonction de densité de perturbation $(D(f)_n)$ *;*

caractérisé par

- la correction de la fonction de densité de perturbation $(D(f)_n)$ en multipliant la fonction de densité de perturbation $(D(f)_n)$ par une fonction de correction pour chaque trame dérivée d'un calcul de corrélation de la densité de puissance de hauteur tonale compensée $(PPX'(f)_n)$ associée au signal d'entrée *(X(t))* d'une trame courante *(n)* et d'une trame précédente indépendante pour obtenir une fonction de densité de perturbation corrigée $(D'(f)_n)$ *;* et
- l'agrégation de la fonction de densité de perturbation $(D'(f)_n)$ en fréquence et dans le temps pour obtenir une mesure (I) de l'intelligibilité du signal de sortie *(Y(t)).*

2. Procédé selon la revendication 1, dans lequel la fonction de correction OrgTempsCorTrame(n) est calculée de la manière suivante :

$$\mathtt{OrgTempsCorTrame(n)} = \mathtt{CorrélationBandedeFréquences}(PPX'(f)_n,\ PPY'(f)_{n-2})$$

3. Procédé selon la revendication 1 ou 2, dans lequel le calcul de corrélation est exécuté sur une plage de domaine de fréquence allant d'une limite de fréquence basse à une limite de fréquence haute, telle que la gamme de 100... 3500 Hz.

4. Procédé selon l'une quelconque des revendications 1 à 3, dans lequel la fonction de correction est limitée à une valeur inférieure ou égale à 1,0, en fonction des règles :

si OrgTempsCorTrame(n) <0,0

CompensationTempsCorrélationTrame = 1,0

ou bien

CompensationTempsCorrélationTrame = 1,0 - (OrgTempsCorTrame(n))^k,

k étant une valeur de puissance prédéterminée.

5. Procédé selon la revendication 4, dans lequel la valeur de puissance prédéterminée est supérieure à 1,0, p. ex. entre 10 et 20.

6. Procédé selon la revendication 4 ou 5, dans lequel la fonction de correction est limitée à une valeur supérieure ou égale à une valeur de limite inférieure, p. ex. 0,4.

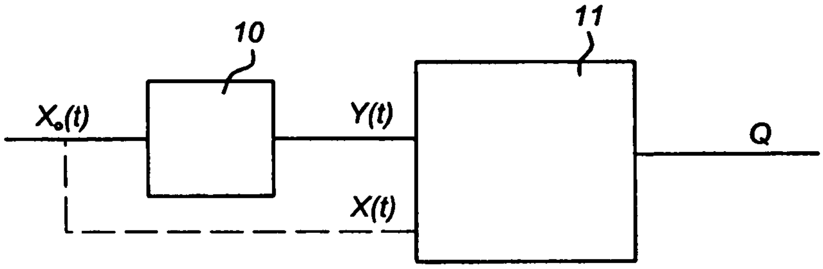
7. Procédé selon l'une quelconque des revendications 1 à 6, dans lequel la fonction de densité de perturbation corrigée ($D'(f)_n$) est agrégée en fréquence en utilisant un facteur de normalisation bas (L_q), le facteur de normalisation bas (L_q) ayant une valeur inférieure ou égale à 2, et agrégé dans le temps en utilisant un facteur de normalisation haut (L_p), le facteur de normalisation haut (L_p) ayant une valeur supérieure ou égale à 6.

8. Procédé selon l'une quelconque des revendications 1 à 6, le procédé comprenant en outre le calcul d'une différence entre deux mesures de score d'intelligibilité (I), dans lequel les mesures de score d'intelligibilité (I) sont calculées en utilisant différents facteurs de normalisation, les facteurs de normalisation étant inférieurs ou égaux à 3.

9. Système de traitement pour mesurer l'intelligibilité d'un signal de sortie dégradé *(Y(t))* depuis un système de transmission audio (10) en réponse à un signal d'entrée de référence *(X(t)),* comprenant un dispositif de mesure (11) connecté au système de transmission audio (10) pour recevoir le signal d'entrée de référence *(X(t))* et le signal de sortie dégradé *(Y(t)),* dans lequel le dispositif de mesure (11) est agencé pour produire une mesure (I) de l'intelligibilité du signal de sortie *(Y(t)),* et exécuter les étapes du procédé selon l'une quelconque des revendications 1 à 8.

10. Produit de programme informatique comprenant un code logiciel exécutable par ordinateur, lequel, quand il est chargé sur un système de traitement, permet au système d'exécuter le procédé selon l'une quelconque des revendications 1 à 8.

Fig 1



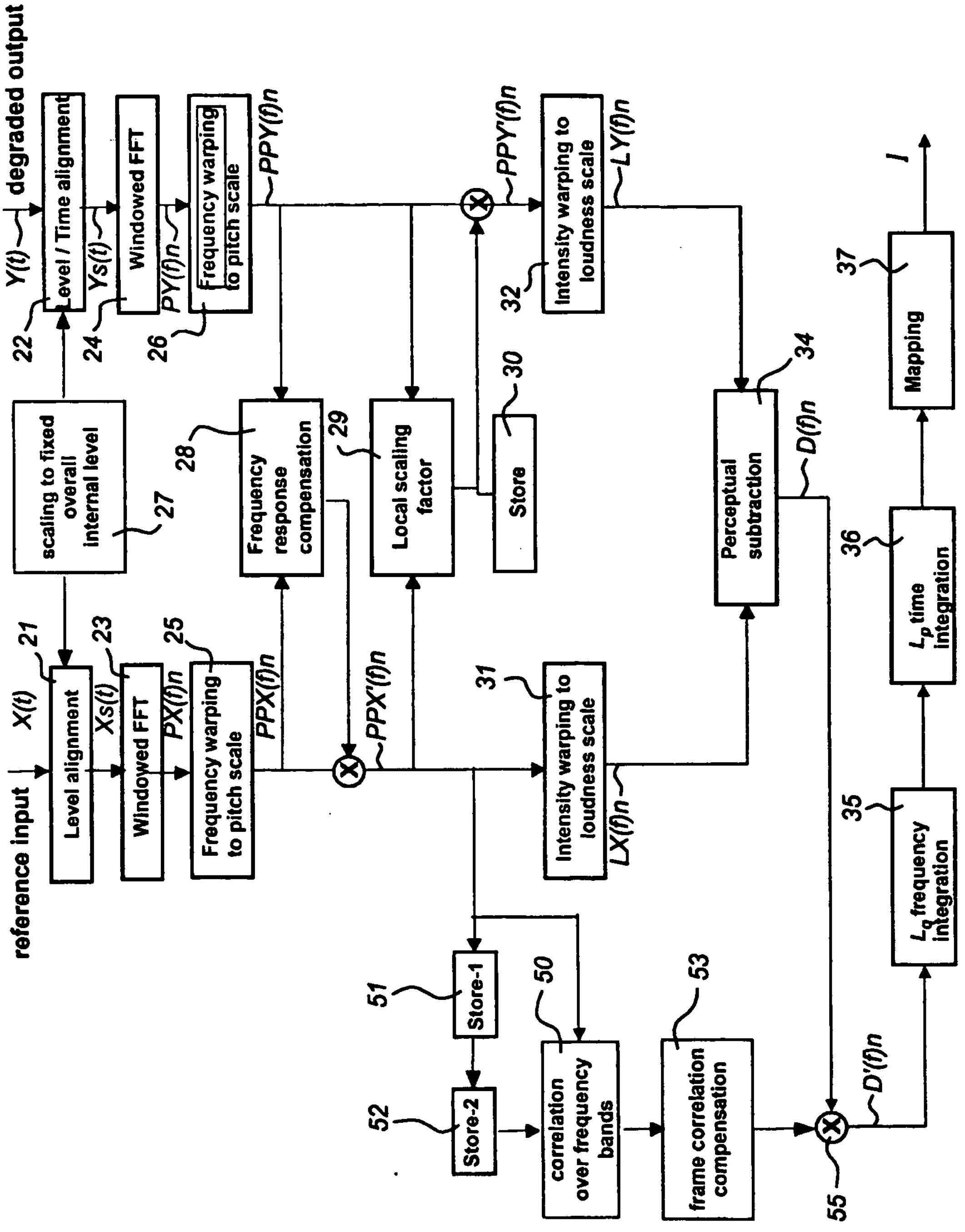

reference input
X(t)
Level alignment
21
Xs(t)
Windowed FFT
23
PX(f)n
Frequency warping to pitch scale
25
PPX(f)n
PPX'(f)n
Intensity warping to loudness scale
31
LX(f)n
degraded output
Y(t)
Level / Time alignment
22
Ys(t)
Windowed FFT
24
PY(f)n
Frequency warping to pitch scale
26
PPY(f)n
PPY'(f)n
Intensity warping to loudness scale
32
LY(f)n
scaling to fixed overall internal level
27
Frequency response compensation
28
Local scaling factor
29
Store
30
Perceptual subtraction
34
D(f)n
Store-1
51
Store-2
52
correlation over frequency bands
50
frame correlation compensation
53
55
D'(f)n
L_q frequency integration
35
L_p time integration
36
Mapping
37
I

Fig 2

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.

Patent documents cited in the description

- EP 402714 A **[0054]**
- EP 0202342 W **[0054]**
- EP 01200945 A **[0054]**
- EP 402736 A **[0054]**
- EP 0205556 W **[0054]**
- EP 01203699 A **[0054]**
- EP 402808 A **[0054]**
- EP 0302058 W **[0054]**
- EP 02075973 A **[0054]**
- EP 42218 A **[0054]**
- EP 402894 A **[0054]**
- EP 03075949 A, Koninklijke PTT Nederland N.V. **[0054]**
- EP 04077601 A **[0054]**

Non-patent literature cited in the description

- **J. Beerends et al.** Perceptual Evaluation of Speech Quality (PESQ), the new ITU standard for end-to-end speech quality assessment, Part II - Psychoacoustic model. *J. Audio Eng. Soc.,* October 2002, vol. 50, 765-778 **[0003]**
- **A. W. RIX ; M. P. HOLLIER ; A. P. HEKSTRA ; J. G. BEERENDS.** PESQ, the new ITU standard for objective measurement of perceived speech quality, Part 1 - Time alignment. *J. Audio Eng. Soc.,* October 2002, vol. 50, 755-764 **[0054]**
- **J. G. Beerends ; A. P. Hekstra ; A. W. Rix ; M. P. Hollier.** Perceptual Evaluation of Speech Quality (PESQ), the new ITU standard for end-to-end speech quality assessment, Part II-Psychoacoustic model. *J. Audio Eng. Soc.,* October 2002, vol. 50, 765-778 **[0054]**
- Perceptual Evaluation Of Speech Quality (PESQ): An Objective Method for End-to-end Speech Quality Assessment of Narrow-band Telephone Networks and Speech Codecs. *International Telecommunication Union,* February 2001, 862 **[0054]**
- Mapping function for transforming P.862 raw result scores to MOS-LQO. *ITU-T Rec.,* November 2003, 862.1 **[0054]**
- Wideband extension to Recommendation P.862 for the assessment of wideband telephone networks and speech codecs. *ITU-T Rec.,* November 2005, 862.2 **[0054]**
- **T. GOLDSTEIN ; J. G. BEERENDS ; H. KLAUS ; C. SCHMIDMER.** Draft ITU-T Recommendation P.AAM, An objective method for end-to-end speech quality assessment of narrow-band telephone networks including acoustic terminal(s). *White contribution COM 12-64 to ITU-T Study Group,* 12 September 2003 **[0054]**
- **H.J.M. STEENEKEN ; T. HOUTGAST.** A physical method for measuring speech-transmission quality. *J. Acoust. Soc. Am.,* January 1980, vol. 67, 318-326 **[0054]**
- **H.J.M. STEENEKEN.** On measuring and predicting speech intelligibility. *PhD University of Amsterdam,* 1992 **[0054]**
- **J. G. BEERENDS ; J. A. STEMERDINK.** A Perceptual Audio Quality Measure based on a psychoacoustic sound representation. *J. Audio Eng. Soc.,* December 1992, vol. 40, 963-978 **[0054]**