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

(11) **EP 2 474 975 A1**

(12) **EUROPEAN PATENT APPLICATION**

(43) Date of publication:
11.07.2012 Bulletin 2012/28

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

(21) Application number: **12000483.3**

(22) Date of filing: **21.05.2010**

(84) Designated Contracting States:
AL AT BE BG CH CY CZ DE DK EE ES FI FR GB GR HR HU IE IS IT LI LT LU LV MC MK MT NL NO PL PT RO SE SI SK SM TR
Designated Extension States:
BA ME RS

(62) Document number(s) of the earlier application(s) in accordance with Art. 76 EPC:
10005327.1 / 2 388 779

(71) Applicant: **SwissQual License AG**
4528 Zuchwil (CH)

(72) Inventor: **Ullmann, Raphel**
1007 Lausanne (CH)

(74) Representative: **Fischer, Britta Ruth et al**
E. BLUM & CO. AG
Vorderberg 11
8044 Zürich (CH)

Remarks:
This application was filed on 26-01-2012 as a divisional application to the application mentioned under INID code 62.

(54) **Method for estimating speech quality**

(57) The invention relates to a method for estimating speech quality, wherein a reference speech signal (301) enters a telecommunication network resulting in a test speech signal (302) and wherein the method comprises the following steps of aligning the reference speech signal (301) and the test speech signal (302) by matching signal parts of the reference speech signal (301) with signal parts of the test speech signal (302), wherein matched signal parts (303, 304) are of similar length in the time domain and have similar intensity summed over their length, and computing and comparing the speech spectra of the reference speech signal (301) and the test speech signal (302) that are aligned, resulting in a difference measure, the difference measure being indicative of the speech quality of the test speech signal.

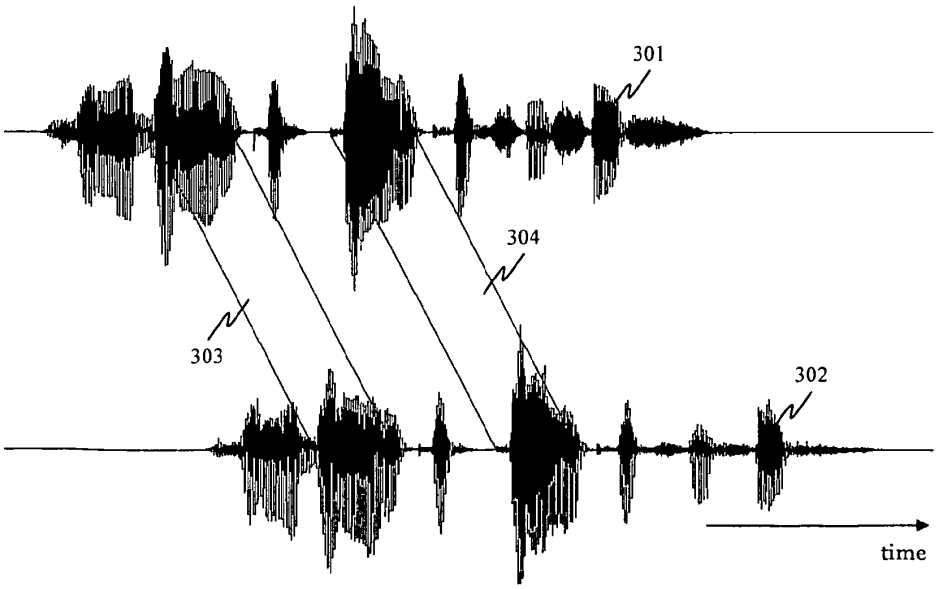

Fig. 7

EP 2 474 975 A1

Description

[0001] The invention relates to a method for estimating speech quality.

[0002] Providers of telecommunication network services have an interest in monitoring the transmission quality of the telecommunication network as perceived by the end-user, in particular with respect to the transmission of speech. For this so-called instrumental or objective methods for speech quality estimation may be used that compare a reference speech signal in form of an undistorted high-quality speech signal, which enters the telecommunication network, with a test speech signal resulting from the reference speech signal, the test speech signal being a speech signal to be tested and analysed, respectively, after transmission via and/or processing by the telecommunication network (including a simulation of a telecommunication network, i.e. a simulated telecommunication network) and possible distortion, wherein the test speech signal is given by the reference speech signal after transmission via and/or processing by the telecommunication network. For comparing the reference speech signal and the test speech signal spectral representations of the respective signals are usually used. The aim of the comparison of the reference speech signal with the test speech signal is the determination of perceptually relevant differences between the reference speech signal and the test speech signal.

[0003] The spectral representations of the reference speech signal and of the test speech signal can be highly influenced by effects that have basically no or little disturbing character for the perception of the end-user such as time differences, e.g. signal delay, or differences in intensity (e.g. power, level or loudness) between the respective speech signals. Usually, such differences are compensated by means of delay/time and intensity alignment procedures before the actual differences between the spectral representations of the reference speech signal and the test speech signal are computed. Both the delay/time and the intensity alignment procedures are not restricted to the compensation of a fixed bias, but can also be applied for time-varying compensation.

[0004] After the compensation of time/delay, intensity and possibly other undesired differences, the remaining differences in the spectral representations of corresponding sections of the reference speech signal and of the test speech signal are used to derive an estimate of their similarity. Typically such similarity estimations are computed for a number of short segments of the reference speech signal and the test speech signal. The similarity estimations computed for the respective segments are then aggregated. The aggregated similarity estimations represent a raw estimation of the overall speech quality of the test signal (i.e. of the network transmission quality) with the raw estimation usually being transformed to a common quality scale such as the known so-called MOS scale (mean opinion score scale) ranging from 1 to 5.

[0005] The similarity estimation can be given by a weighted spectral difference or by a kind of spectrum correlation between the reference speech signal and the test speech signal. The spectral representation of a signal segment is often derived from the common short-term Fourier transform that has been further transformed to a perceptual representation of its frequency content, e.g. by applying common loudness models as described in Paulus, E. and Zwicker, E., "Programme zur automatischen Bestimmung der Lautheit aus Terzpegeln oder Frequenzgruppenpegeln", Acustica, Vol. 27, No. 5, 1972.

[0006] Almost all methods for speech quality estimation that are commercially available today follow the outlined approach that is schematically depicted in Figure 1. A reference speech signal 1 enters e.g. by means of a mobile phone 2 a telecommunication network 3 resulting in a usually degraded test speech signal 4 received e.g. by a mobile phone 5 or tapped/scanned after receipt e.g. by the mobile phone 5, that shall be perceived by an end-user 6. Box 7 illustrates the method for speech quality estimation. First time/delay and intensity differences are compensated by alignment procedures (box 8). Then the spectral representations of the aligned speech signals 1, 4 are computed and compared to give similarity estimations (box 9), wherein the computation and the comparison is typically performed for short segments of the respective signals during their entire duration (illustrated by arrow 10). From the similarity estimations the speech quality of the test speech signal is estimated in box 11.

[0007] Corresponding known methods are described in Beerends, J.G., Hekstra, A.P., Rix, A.W., Hollier, M.P., "Perceptual Evaluation of Speech Quality (PESQ), the new ITU standard for end-to-end speech quality assessment. Part I - Time alignment", J. Audio. Eng. Soc., Vol. 50, No. 10, October 2002, Beerends, J.G., Hekstra, A.P., Rix, A.W., Hollier, M.P., "Peceptial 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, No. 10, October 2002, 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, December 1994, and ITU-T, Study Group 12, "TOSQA - Telecommunication objective speech quality assessment", COM12-34-E, Geneva, December 1997.

[0008] The differences between the known methods lie mainly in the implementation details of the steps corresponding to the boxes 8, 9 and 10 in Figure 1, as well as in the way the spectral representations of the respective signals are transformed to a perceptual representation of their frequency content. They also use different strategies for post-processing and weighing of the raw estimations (i.e. the above-mentioned similarity estimations). The aim of each known method is to achieve a high prediction accuracy of the computed overall speech quality when compared to speech quality values obtained in listening tests with human participants. Usually, meth-

ods for speech quality estimation try to predict the mean opinion score (MOS) obtained in so-called absolute category rating auditory experiments with human participants (ITU-T Recommendation P.800, "Methods for subjective determination of transmission quality", Geneva, 1996).

[0009] In order to estimate the speech quality of a test speech signal it is desirable to determine which frequency bands in the test speech spectrum Y(f) have been amplified or attenuated, respectively, when compared to the undistorted reference speech spectrum X(f). According to a known, rather simple method the intensity difference between the test speech spectrum Y(f) and the reference speech spectrum X(f) is determined by calculating the difference $D_1(f) = Y(f) - X(f)$ for each frequency band f. However, changes to the speech spectrum intensity that are constant over all frequency bands, as e.g. caused by corresponding signal attenuation properties of the transmission channel, usually contribute to the perceived speech quality only to a limited extent. Rather, changes/modifications in the relative intensity of the frequency bands within the speech spectrum have been found to have a more significant effect on the perceived speech quality.

[0010] Figure 2 shows an example of a reference speech spectrum and a test speech spectrum, wherein the test speech spectrum is uniformly attenuated over all frequency bands when compared with the reference speech spectrum. The calculation of the difference $D_1(f)$ may yield a large absolute intensity difference despite the limited impairment of the perceived speech quality. Figure 3 shows a further example of a reference speech spectrum and a test speech spectrum, wherein the test speech spectrum Y(f) differs from the reference speech spectrum X(f) only in a single frequency band f_i. For this example the calculation of the difference $D_1(f)$ yields the desired measure of the perceived intensity difference, as the only non-zero result for $D_1(f)$ is obtained for the frequency band f being equal to f_i.

[0011] Alternatively, it is known to compute the intensity difference by using a scaled test speech spectrum $Y_2(f) = c \cdot Y(f)$ with c being a scaling factor and given by c = average(X(f))/average(Y(f)) such that the average intensities of both spectra are aligned. The intensity difference $D_2(f)$ is then given by $D_2(f) = Y_2(f) - X(f)$. For the example depicted in Figure 2 the computed intensity difference $D_2(f)$ is equal to zero for all frequency bands f, thus yielding a result that is closer to the actually perceived speech quality impairment. For the example depicted in Figure 3 the calculation of $D_2(f)$ however results in non-zero values for all frequency bands f, which does not correspond to the perceived intensity difference as $D_2(f)$ is not only non-zero for the frequency band f_i.

[0012] For time/delay compensation, known approaches are not just capable of computing the overall time difference between the reference speech signal and the test speech signal in time domain, but they can determine the time differences between individual parts of the respective signals. For this corresponding parts of the reference speech signal and of the test speech signal are matched. The signal parts are matched and their respective time differences are typically computed in the order of the temporal/chronological occurrence of the signal parts in the respective speech signals, i.e. signal parts occurring at the end of the respective signals are matched after signal parts occurring at the beginning of the respective speech signals have already been matched.

[0013] Figure 6 shows a corresponding example with a reference speech signal 201 and a test speech signal 202 in time domain. The test speech signal 202 exhibits a positive time difference, i.e. it starts later in time, when compared to the reference speech signal 201. The same applies to its signal parts when compared with corresponding signal parts of the reference speech signal 201. The known matching procedure starts at the beginning of the signals 201, 202 and progresses monotonously in time, yielding e.g. the matched signal parts 203 and 204. However, signal parts of the reference speech signal 201 following the match 204 can be matched with any signal part of the test speech signal 202 that lies chronologically after the match 204. The already matched signal parts 203 and 204 thus limit the number of possible matches by not taking into account signal parts chronologically occurring before the signal part that is currently matched. This approach can therefore lead to incorrect matching as illustrated by the erroneous match 205. Such a matching procedure that starts at the beginning of the speech signals 201 and 202 and progresses monotonously in time may only lead to a limited extent to correct matching of later occurring signal parts. Consider in this respect also the example of a test speech signal whose beginning has been muted. A miscalculation in known approaches for time/delay compensation may lead to the non-muted beginning of the reference speech signal being matched with an intact, non-muted part of the test speech signal that shares some similarities with the beginning of the reference speech signal that has been muted in the test speech signal, but occurs chronologically later in the test speech signal. As a consequence, the part of the reference speech signal, that actually corresponds to the aforementioned intact part of the test speech signal that wrongly has been matched to the beginning of the reference speech signal, can only be matched with signal parts of the test speech signal occurring after this already matched intact part by means of the above-described known approach (cf. Figure 6). This follows from the fact that each signal part can only be matched once and from the matching being typically performed such that the temporal order of matched signal parts in either signal is preserved. Therefore, the occurrence of one incorrect matching may bias or deteriorate, respectively, the matching of signal parts occurring chronologically later in the respective speech signals.

[0014] One of the typical problems of speech signal transmission is the interruption or loss of speech. Known approaches rate the portions of the test speech signal

with missed speech by comparing the test speech signal with the reference speech signal and measuring the amount of missed speech intensity, wherein the amount of missed speech intensity can be computed from perceptual representations of the speech signals such as loudness spectra. Hence, with known approaches the amount of missed speech is related to the part of the reference speech signal that has actually been missed. However, this approach might be disadvantageous as a human listener who listens to the test speech signal does not rate missed speech in such manner. As the human listener has no knowledge of the reference speech signal, he has no possibility to compare the test speech signal with the reference speech signal, and he hence has no knowledge of what is actually missing. As a consequence, the actually perceived distortion that is caused by an interruption or loss of speech is rather related to the knowledge and expectations of the human listener formed on the basis of the received portions of the test speech signal.

[0015] It is an object of the invention to provide a method for estimating speech quality, by which a speech quality estimate can be obtained which is close to the speech quality perceived by an end-user (also referred to as human listener), in particular for the examples of the reference speech spectrum and the test speech spectrum depicted in Figures 2 and 3. It is a further object of the invention to provide a method for estimating speech quality with improved time/delay compensation. It is a still further object of the invention to provide a method for estimating speech quality, by which interruptions or loss of speech, respectively, can be handled in a satisfactory manner. It is a still further object of the invention to provide a method for estimating speech quality that follows the basic approach depicted in and described with reference to Figure 1.

[0016] In order to implement these and still further objects of the invention, which will become more readily apparent as the description proceeds, a method for estimating speech quality is provided, wherein a reference speech signal enters a telecommunication network, in particular a mobile network, resulting in a test speech signal, and that comprises the steps of aligning the reference speech signal and the test speech signal by matching signal parts of the reference speech signal with signal parts of the test speech signal, wherein matched signal parts of the respective signals are of similar length in time domain and have similar intensity summed over their length or relative to their length, and of computing and comparing the speech spectra of the aligned reference speech signal and the aligned test speech signal, the comparison resulting in a difference measure that is indicative of the speech quality of the test speech signal.

[0017] For matching signal parts of the reference speech signal with corresponding signal parts of the test speech signal, first the one or more signal parts of the reference speech signal that have the highest intensity summed over their length or relative to their length, respectively, are matched with signal parts of the test speech signal. Then the matching continues by matching signal parts of the reference speech signal with signal parts of the test speech signal, whereby for the signal parts of the reference speech signal to be matched the intensity summed over the length or relative to the length, respectively, decreases for each subsequent match.

[0018] A performance measure is preferably computed for each pair of matched signal parts. The performance measure is in particular given by the maximum of the cross-correlation of the matched signal parts that is normalized by the signal powers of the matched signal parts. If the performance measure of a pair of matched signal parts lies beneath a pre-set threshold value, then the pair of matched signal parts is preferentially deemed to have insufficient performance. Alternatively, a pair of matched signal parts is deemed to have insufficient performance if its performance measure is significantly lower than the performance measures of other pairs of matched signal parts. If a pair of matched signal parts is deemed to have insufficient performance, then its signal parts are preferably re-matched, i.e. matched again with other signal parts of the respective signal. In particular, a pair of matched signal parts that is deemed to have insufficient performance may be un-matched, i.e. the corresponding respective signal parts may be made available again for matching with other signal parts that have not yet been matched. Re-matching of the now again unmatched reference speech signal part may be performed after further other reference speech signal parts have been matched. Hence, employing the performance measure may result in a matching order of the reference speech signal parts that differs from a matching order given by the reference speech signal parts arranged in accordance to their respective intensity summed over or relative to their respective length with decreasing intensity.

[0019] With the method according to the invention incorrect matching of signal parts of the reference speech signal with signal parts of the test speech signal can advantageously be reduced if not even avoided.

[0020] The method of the invention may also comprise the further steps of identifying a number of perceptually dominant frequency sub-bands in one of the reference speech spectrum and the test speech spectrum, with the reference speech signal having a reference speech spectrum and the test speech signal having a test speech spectrum, computing an intensity scaling factor for each identified sub-band by minimizing a measure of the intensity difference between those parts of the reference speech spectrum and the test speech spectrum that correspond to the respective sub-band, multiplying the test speech spectrum with each intensity scaling factor thus generating a number of scaled test speech spectra, selecting one scaled test speech spectrum, and computing the difference between the selected scaled test speech spectrum and the reference speech spectrum. This difference is indicative of the speech quality of the test speech signal. The measure of the intensity difference

is preferably given by the squared intensity difference or the global maximum of the intensity difference between those parts of the reference speech spectrum and of the test speech spectrum that correspond to the respective sub-band.

[0021] The number of perceptually dominant sub-bands of one of the reference speech spectrum and the test speech spectrum is preferably identified by determining the local maxima in a perceptual representation of the respective spectrum and by selecting a predetermined frequency range around each local maximum, wherein the predetermined frequency range is preferentially determined by the local minima bordering the respective local maximum, with one local minimum on each side (in frequency domain) of the respective local maximum. In particular the predetermined frequency range shall be smaller or equal to 4 Bark.

[0022] The perceptual representation of the respective spectrum is preferably obtained by transforming the respective spectrum to a loudness spectrum as e.g. defined in Paulus, E. and Zwicker, E., "Programme zur automatischen Bestimmung der Lautheit aus Terzpegeln oder Frequenzgruppenpegeln", Acustica, Vol. 27, No. 5, 1972.

[0023] The selected scaled test speech spectrum is preferably given by the scaled test speech spectrum yielding the lowest measure of the intensity difference between the reference speech spectrum and a scaled test speech spectrum with the intensity difference being computed for each scaled test speech spectrum. The measure of the intensity difference is preferably given by the squared intensity difference or alternatively the global maximum of the intensity difference between the reference speech spectrum and a respective scaled test speech spectrum.

[0024] Through appropriate scaling of the test speech spectrum a difference between the reference speech spectrum and the test speech spectrum can be computed that is close to human perception, that in particular basically does not take into account amplifications and attenuations, respectively, that are constant over all frequency bands, but that places emphasis on modifications in the relative intensity of single frequency bands that contribute to a qualitative impairment that would be perceived by a human listener.

[0025] The method of the invention may also comprise the further steps of for at least one missed or interrupted signal part in the test speech signal computing the signal intensities of the signal parts of the test speech signal that are adjacent to the missed or interrupted signal part, deriving an expected signal intensity for the at least one missed or interrupted signal part from the computed signal intensities of the adjacent signal parts, computing a measure of the perceived distortion by comparing the actual intensity of the at least one missed or interrupted signal part in the test speech signal with the derived expected intensity for the at least one missed or interrupted signal part, computing a measure of the actual distortion by comparing the reference speech signal with the test speech signal, and combining the measure of the perceived distortion with the measure of the actual distortion to generated a combined measure of distortion that is indicative of the speech quality of the test speech signal.

[0026] The expected signal intensity of the at least one missed or interrupted signal part in the test speech signal is preferably derived from the computed signal intensities of the adjacent signal parts of the test speech signal by means of interpolation, in particular by means of linear interpolation and/or spline interpolation.

[0027] By considering the signal intensities of signal parts of the test speech signal that are adjacent to an interruption or speech loss it can be better assessed how such a distortion is actually perceived. The method according to this still further aspect of the invention is hence advantageously perceptually motivated.

[0028] The method of the invention can advantageously be combined with existing methods for speech quality estimation that in particular have the structure depicted in and described with respect to Figure 1 to improve and extend the existing methods.

[0029] Further advantageous features and applications of the invention can be found in the dependent claims, as well as in the following description of the drawings illustrating the invention. In the drawings like reference signs designate the same or similar parts throughout the several figures of which:

Fig. 1 shows a block diagram illustrating the basic steps for speech quality estimation,
Fig. 2 shows a first example of a reference speech spectrum and of a test speech spectrum,
Fig. 3 shows a second example of a reference speech spectrum and of a test speech spectrum,
Fig. 4 shows a flow chart of an embodiment of a method for estimating speech quality,
Fig. 5 shows a diagram illustrating the embodiment of the method for estimating speech quality,
Fig. 6 shows a diagram illustrating a method for speech quality estimation according to the state of the art,
Fig. 7 shows a diagram illustrating an embodiment of the method of the invention,
Fig. 8 shows a flow chart of a further embodiment of a method for estimating speech quality, and
Fig. 9 shows a diagram illustrating a further embodiment of a method for estimating speech quality.

Figures 1, 2, 3, and 6 have been described in the introductory part of the description and reference is made thereto.

[0030] Figure 4 shows a flow chart of a first embodiment of a method for estimating speech quality. In a first step 20 of this embodiment a certain number N of perceptually dominant frequency sub-bands $b_{1...N}$ is identified in and selected from one of the reference speech spectrum X(f) and the test speech spectrum Y(f), for example from the undistorted reference speech spectrum

X(f), or from both spectra. The reference speech spectrum X(f) and the test speech spectrum Y(f) represent exemplary speech spectra of a reference speech signal and a test speech signal, respectively, which can both comprise several speech spectra. The sub-bands bi are given by bi = $[f_j...f_k]$, where f represents the frequency and where i, j, k, N are integers with $k \geq j$ and i = 1...N. Using only perceptually dominant frequency sub-bands leads to the first embodiment of the method for estimating speech quality being perceptually motivated.

[0031] For the identification of the perceptually dominant frequency sub-bands the respective spectrum may be transformed to a perceptual representation of the respective spectrum, the perceptual representation corresponding to the frequency content that is actually received by a human auditory system. Then the local maxima of the perceptual representation are identified and a predetermined range of frequencies around each local maximum gives the perceptually dominant frequency sub-bands. The limiting values of each predetermined range are preferentially given by the local minima adjacent to the respective local maximum with the condition that the entire range is in particular smaller or equal to 4 Bark. The loudness spectrum is one example of a perceptual representation that has been found to represent to a high degree the human subjective response to auditory stimuli.

[0032] In a second step 21 of this first embodiment of the method for estimating speech quality, an intensity scaling factor c_i is preferably computed for each identified sub-band b_i. The respective intensity scaling factor c_i is computed such that the squared intensity difference $|X(b_i)\text{-}ci\cdot Y(b_i)|^2$ between both spectra inside the respective identified sub-band bi is minimized.

[0033] In the following step 22 of this embodiment the test speech spectrum Y(f) is multiplied with each intensity scaling factor c_i, thereby generating a number N of scaled test speech spectra given by $Y_i(f) = c_i \cdot Y(f)$.

[0034] Then in step 23 one scaled test speech spectrum $Y_{sel}(f)$ of the generated scaled test speech spectra $Y_i(f)$ is selected from the generated N scaled test speech spectra $Y_i(f)$. The selection of the scaled test speech spectrum $Y_{sel}(f)$ may be achieved by first computing the total squared intensity difference between the reference speech spectrum X(f) and each of the generated N scaled test speech spectra $Y_i(f)$ over all frequency bands in the spectrum and by then selecting that particular scaled test speech spectrum $Y_{sel}(f)$ that yields the lowest total squared intensity difference such that the index sel is given by $sel = argmin_i(sum(|X(f) - Y_i(f)|)^2))$.

[0035] Finally, in step 24 the difference between the selected scaled test speech spectrum $Y_{sel}(f)$ and the reference speech spectrum X(f) is computed in form of the spectral difference function D(f) that is given by $D(f) = Y_{sel}(f) - X(f)$. The spectral difference function D(f) contains non-zero values for frequency bands that have been amplified and attenuated, respectively, when compared with the reference speech spectrum X(f). Positive values of D(f) correspond to amplified spectrum portions and negative values of D(f) correspond to attenuated spectrum portions. For frequencies f lying inside the sub-band bi that corresponds to the scaling factor c_i of the selected scaled test spectrum $Y_{sel}(f)$, the spectral difference function D(f) normally contains small absolute values. The spectral difference function D(f) constitutes an estimate of the speech quality.

[0036] In case of an amplification or attenuation of the reference test spectrum intensities that is uniform over all frequency bands as depicted in Figure 2 for the case of an attenuation, the computation of the spectral difference function D(f), that includes the computation and selection of a scaled test speech spectrum, fully compensates the difference between the reference speech spectrum X(f) and the test speech spectrum Y(f) yielding a spectral difference function D(f) that is zero at all frequencies f.

[0037] In the second example depicted in Figure 3 the test speech spectrum Y(f) differs from the reference speech spectrum X(f) only in a single frequency band f_i. In this case the values of the computed spectral difference function D(f) depend on whether the frequency band f_i is part of the selected sub-band b_{sel}, i.e. the sub-band bi whose intensity scaling factor c_i is the scaling factor of the selected scaled test speech spectrum $Y_{sel}(f)$. If fi lies outside the selected sub-band b_{sel}, then the calculated scaling factor c is equal to 1 and the spectral difference function D(f) has non-zero values only for f being equal to f_i. If however the frequency band f_i is part of the selected sub-band b_{sel}, then the value of the scaling factor c depends on the modified intensity at the frequency f_i (modified in comparison to the reference speech intensity) and the selected scaled test speech spectrum $Y_{sel}(f)$ differs from the reference speech spectrum X(f) at frequencies other than f_i. The spectral difference function D(f) hence has a large number of non-zero values, thereby reflecting the expected larger impact of a modification of intensities at a frequency band that belongs to a perceptually dominant sub-band.

[0038] Hence, the first embodiment of the method for estimating speech quality computes a difference that is indicative of the speech quality of the test speech signal in form of the spectral difference function D(f) that provides better approximations of the perceptions of speech spectrum intensity changes by a human listener when compared with existing methods.

[0039] Figure 5 illustrates a possible application of the first embodiment of the method for estimating speech quality. An example of a perceptual representation of a reference speech spectrum 101 is shown along with an example of a perceptual representation of a test speech spectrum 102. Compared to the perceptual representation of the reference speech spectrum 101 the perceptual representation of the test speech spectrum 102 features an amplification of intensities at lower frequencies, as well as a limitation of the bandwidth leading to a strong attenuation of the intensities at higher frequencies. With-

out knowledge of the original, undistorted reference speech signal, the slight amplification at lower frequencies is of rather limited influence on the perception of speech quality by a human listener. However, the change in the relative intensities of the various frequency bands (when compared to each other) within the perceptual representation of the test speech spectrum 102, as well as the limitation of the bandwidth have a much higher impact on the perceived speech quality.

[0040] For the example depicted in Figure 5 the perceptually dominant frequency sub-bands are identified in the perceptual representation of the reference speech spectrum 101 as described above, i.e. by determining the local maxima and selecting a predetermined frequency range around each local maximum. Highlighted area 104 corresponds to one such perceptually dominant sub-band. Each identified perceptually dominant sub-band gives rise to an intensity scaling factor and a correspondingly scaled test speech spectrum. The dotted curve 103 in Figure 3 represents a perceptual representation of a scaled test speech spectrum that has been scaled with the intensity scaling factor associated with the sub-band 104.

[0041] Comparing the perceptual representation of the reference speech spectrum 101 with the perceptual representation of the scaled test speech spectrum 103 it can be seen from Figure 5 that the slight intensity difference at low frequencies between the perceptual representation of the reference speech spectrum 101 and the perceptual representation of the test speech spectrum 102 is strongly reduced and in this particular case even equal to zero. The modification of the relative intensities of the frequency bands within the perceptual representation of the test speech spectrum 102 when compared with the perceptual representation of the reference speech spectrum 101 is still present in the perceptual representation of the scaled test speech spectrum 103 and can be measured as negative difference (i.e. attenuated spectrum portions) at middle frequencies between the perceptual representation of the reference speech spectrum 101 and the perceptual representation of the scaled test speech spectrum 103 (in Figure 5: basically that portion of curve 103 that does not overlap with any other curve). Also the limitation of the bandwidth is still present in the perceptual representation of the scaled test speech spectrum 103, leading to a large negative difference at higher frequencies when compared with the perceptual presentation of the reference speech spectrum 101. This is in line with the change in the relative intensities of frequency bands and with a limitation of the bandwidth at higher frequencies generally having a large impact on the perceived speech quality.

[0042] As described in the introductory part of the description, with known approaches for time/delay compensation corresponding signal parts of the reference speech signal and the test speech signal are matched and their time difference is computed in the order of the temporal/chronological occurrence of the respective signal parts in the speech signals, starting at the beginning of the speech signals. This may, however, result in erroneous matches for subsequent speech signal parts if a miscalculation or erroneous match occurs at the beginning of the speech signals. This has been described in detail above in connection with Figure 6.

[0043] An embodiment of the method of the invention avoids this disadvantage in that it attempts to first match signal parts of the reference speech signal with corresponding signal parts of the test speech signal, that are least likely to result in erroneous matches. This is achieved by first starting to match the one or more parts of the reference speech signal with the highest intensity summed over their length, e.g. the parts of the reference speech signal with the highest signal energy or loudness summed over their length. For the matching cross-correlation may be employed. Instead of the highest intensity summed over the respective length the highest intensity relative to the respective length may be used, and hence the highest signal energy or loudness relative to the respective length. The parts of the reference speech signal identified to have the highest intensity summed over their length (or relative to their length, respectively), are matched according to the order of decreasing summed intensity (or decreasing relative intensity, respectively). Degradations in the test speech signal such as introduced e.g. by packet loss concealment routines in packetized transmission systems often result in decreased signal energies of correspondingly degraded signal parts in the test speech signal. High-energy signal parts of the reference signal are therefore more likely to be still present with sufficiently high energy in the test speech signal in comparison to low energy parts. The length of signal parts to be matched can be in the range of 0.05 to 0.5 seconds.

[0044] After first matching signal parts of the reference speech signal with the highest intensity summed over their length (or relative to their length, respectively), the embodiment of the method of the invention attempts to match signal parts of the reference signal with decreasing intensity summed over the length (or relative to their length, respectively), i.e. it attempts to match signal parts in order of decreasing expected match accuracy rather than monotonously progressing from the beginning of the reference speech signal to the end of the reference speech signal. Such the possibility of erroneous matching decreases with each further matched signal part of the reference speech signal, since the remaining amount of matchable signal parts in the test speech signal is limited by the amount of already matched signal parts, normally surrounding the signal parts still to be matched in the time domain.

[0045] Before matching the signal parts of the reference speech signal, the reference speech signal and the test speech signal are preferably pre-filtered by a bandpass filter to filter out irrelevant signal parts such as background noise. The bandpass filter is preferably configured such that it passes frequencies within the audio

band, in particular in the range of 700 Hz to 3000 Hz, and rejects or at least attenuates frequencies outside the thus defined range.

[0046] The reference speech signal and the test speech signal are further preferably thresholded, i.e. limited by a predefined threshold, and normalized with respect to their respective signal energies/signal powers to compensate for differences between corresponding signal parts of the reference speech signal and the test speech signal that are considered irrelevant. Such irrelevant differences may, for example, be caused by varying gain properties of the transmission channel/telecommunication network in question. The computational operations that are performed for the thresholding and normalization are preferably configured and performed such that a sliding window of preferably 26.625 ms length is moved over the entire length of both speech signals in time domain, that for each speech signal the average signal power within the sliding window is computed while the sliding window is moved over the respective speech signal, and that the average signal power within each sliding window is re-scaled to either a first predefined value if it exceeds a pre-set threshold value, or otherwise is set to a second predefined value. This pre-set threshold value is preferentially set equal to (S + 3*N)/4 with S being the average signal level of the speech content within the respective speech signal and N being the signal level of the background noise in the respective speech signal. The value for S may, for example, be computed as described in ITU-T Recommendation P.56 "Objective measurement of active speech level", Geneva, 1993. The second predefined value is chosen smaller than the first predefined value. The second predefined value may e.g. be equal to 0.

[0047] For computing the intensity summed over the length of signal parts of the respective speech signals, the respective intensities are preferably compared with a second pre-set threshold value and only those intensities are taken into account and summed up that exceed this second pre-set threshold value. The second pre-set threshold value lies preferentially slightly above the above-mentioned first threshold value (S + 3*N)/4. The second pre-set threshold value is preferably given by 0.4*S + 0.6*N with S and N as defined in the last paragraph.

[0048] Figure 7 shows a diagram illustrating the embodiment of the method of the invention. In the diagram a reference speech signal 301 and a test speech signal 302 are shown in the time domain. By first matching those signal parts of the reference speech signal 301 with corresponding signal parts of the test speech signal 302 that are least likely to result in erroneous matches, i.e. by first matching those signal parts with the highest intensity summed over their length (or relative to their length, respectively), the speech signals 301 and 302 are subdivided into smaller sections. In Figure 7 these sections are given by the respective speech signals 301 and 302 without the already matched signal parts 303 and 304. Signal parts within the remaining sections of the reference speech signal 301 can only be matched with signal parts of the test speech signal 302 that occur in the corresponding section of the test speech signal 302, with the temporal locations of the already matched signal parts surrounding or limiting, respectively, the sections.

[0049] In the example shown in Figure 7 the signal part of the reference speech signal 301 between the matches 303 and 304 can only be matched with a corresponding signal part of the test speech signal 302, i.e. with a signal part of the test speech signal 302 that lies between the signal parts of the test speech signal 302 of the matches 303 and 304 in the time domain. The embodiment of the method of the invention thus reduces the possibility of incorrect matching by subdividing the reference speech signal and the test speech signal into smaller sections, the sections being separated by already performed matches.

[0050] Preferably a performance measure (also called performance metric) is computed for each matched pair 303, 304 of signal parts. The performance measure may for example be given by the maximum of the waveform cross-correlation of the matched signal parts of the reference speech signal and the test speech signal, the waveform cross-correlation being normalized by the signal powers of the respective signal parts. A decision unit may be provided to assess the performance of each pair of matched signal parts by evaluating their associated performance measure. The decision unit evaluates if the performance measure is equal to or exceeds a pre-set threshold value. If the value of the performance measure is neither equal to nor exceeds the pre-set threshold value then the decision unit interprets this finding as the pair of matched signal parts having insufficient performance, i.e. as the match being poor.

[0051] The decision unit may also compare the performance measure for a particular pair of matched signal parts with the performance measures computed for other pairs of matched signal parts or with the average value of the performance measures computed for other pairs of matched signal parts, respectively. If the performance measure of the particular pair of matched signals is significantly lower than the performance measures (or the average of the performance measures) of the other pairs of matched signal parts, i.e. if the difference between the performance measure of the particular pair of matched signals and the performance measures (or the average of the performance measures) of the other pairs of matched signal parts exceeds a pre-defined threshold value, then the decision unit may assess the particular pair of matched signal parts as having insufficient performance.

[0052] If a pair of matched signal parts is assessed as having insufficient performance, then the decision unit may reject the particular pair of matched signal parts and skip those signal parts, so that the signal parts may be used for later matching, i.e. may be re-matched later. The matching is then preferably first continued for differ-

ent signal parts, thus subdividing the reference speech signal and the test speech signal into smaller sections. Matching of the skipped signal parts is then preferably reattempted when the possibility of erroneous matching has been further reduced by further subdivision of the reference speech signal and the test speech signal, or when no other unmatched signal parts of the reference signal are left for matching.

[0053] When transmitting speech signals via a telecommunication network distortions may occur due to interruptions of the speech signal or missed speech (missed parts of the speech signal) caused for example by a temporary dead spot within the telecommunication network. Common approaches calculate the amount of distortion caused by such an interruption or loss of speech based on the signal intensity (e.g. the power, level or loudness) that is missing or decreased in the test speech signal when compared to the corresponding signal part(s) in the reference speech signal. However, these common approaches do not take into account that a human listener who listens to the test speech signal has no knowledge of the reference speech signal as such and thus does not know how much signal intensity is actually missing.

[0054] According to a second embodiment of a method for estimating speech quality the test speech signal is analysed shortly before and shortly after the location of an occurrence of an interruption or a speech loss, i.e. the analysing takes place at instances (i.e. signal parts) in the test speech signal that are known to a human listener. It is expected that low signal intensities (e.g. power, level or loudness) in these signal parts of the test speech signal lead to a relatively low perceived distortion for a human listener. Even though the actually missed or interrupted signal part may be of higher signal intensity than the remaining surrounding signal parts, it is assumed that a human listener does not perceive the interruption or speech loss as strong since he does not expect the signal intensity to be high, his expectation being based on the lower signal intensity of the surrounding signal parts in the test speech signal.

[0055] Figure 9 depicts an example of a reference speech signal 401 and a test speech signal 402 in the time domain, wherein a signal part 403 with high signal intensity is lost during transmission and thus missing in the test speech signal 402. The corresponding signal part 404 in the test speech signal has in comparison extremely low signal intensity.

[0056] Figure 8 depicts a flow chart of the second embodiment of the method for estimating speech quality. In a first step 30 of the second embodiment of the method for estimating speech quality the signal intensities of the signal parts of the test speech signal are computed that lie adjacent to the missed or interrupted signal part, i.e. that surround the interruption or speech loss. In a next step 31 the expected signal intensity at the location of the interruption or speech loss in the test speech signal is computed. This expected signal intensity is derived from the computed signal intensities of the adjacent signal parts that have been computed in step 30. The expected signal intensity at the interruption or speech loss may be derived from the computed signal intensities of the adjacent signal parts by means of interpolation, in particular by means of linear and/or spline interpolation.

[0057] In a next step 32 a measure of the perceived distortion is computed by comparing a test speech signal, in which the interruption or loss has been replaced by the derived expected signal intensity of the missed or interrupted signal part, with the actual test speech signal. In particular the measure of the perceived distortion is computed by comparing the actual intensity of the at least one missed or interrupted signal part in the test speech signal with the derived expected intensity for the at least one missed or interrupted signal part. The computed measure of the perceived distortion lies preferable in the range of 0 to 1. In step 33 a measure of the actual distortion is computed by comparing the reference speech signal with the actual test speech signal. The order of steps 32 and 33 can be interchanged. Steps 32 and 33 may also be performed concurrently.

[0058] Finally, in step 34 the computed measure of the perceived distortion is combined with the computed measure of the actual distortion yielding a combined measure of distortion that may be used to assess the speech quality impairment caused by the interruption or speech loss. For combining the measure of the perceived distortion with the measure of the actual distortion, the measure of the perceived distortion may be multiplied with the measure of the actual distortion to compute the combined measure of distortion. Additionally or alternatively, the combined measure of distortion may be given by the measure of the actual distortion limited to the measure of the perceived distortion if the measure of the actual distortion exceeds the measure of the perceived distortion. Still additionally or alternatively, the combined measure of distortion may be given by the measure of the actual distortion exponentially weighted by the measure of the perceived distortion, i.e. by an exponentiation with the measure of the actual distortion being the base and the measure of the perceived distortion being the exponent. Still additionally or alternatively, the combined measure of distortion may be given by the difference (computed through subtraction) between the measure of the perceived distortion and the measure of the actual distortion. The above-mentioned ways for computing the combined measure of distortion may be combined in any perceivable ways.

[0059] Through the computation of expected signal intensities of missed or interrupted signal parts from signal intensities of adjacent signal parts in the test speech signal, it can be better assessed by the second embodiment of the method for estimating speech quality how such a distortion is actually perceived by a human listener.

[0060] The first and the second embodiments of the method for estimating speech quality and the embodiment of the method of the invention may be combined in

various combinations, i.e. the first embodiment of the method for estimating speech quality may be combined with the second embodiment of the method for estimating speech quality and/or the embodiment of the method of the invention, the embodiment of the invention may be combined with the first and/or the second embodiment of the method for estimating speech quality, and the second embodiment of the method for estimating speech quality may be combined with the first embodiment for estimating speech quality and/or the embodiment of the method of the invention.

Claims

1. A method for estimating speech quality,
wherein a reference speech signal (301) enters a telecommunication network resulting in a test speech signal (302), the method comprising the following steps:

    - aligning the reference speech signal (301) and the test speech signal (302) by matching signal parts of the reference speech signal (301) with signal parts of the test speech signal (302), wherein matched signal parts (303, 304) are of similar length in the time domain and have similar intensity summed over their length, and
    - computing and comparing the speech spectra of the reference speech signal (301) and the test speech signal (302) that are aligned, resulting in a difference measure, the difference measure being indicative of the speech quality of the test speech signal,
    wherein for the matching of signal parts of the reference speech signal (301) with signal parts of the test speech signal (302), first the one or more signal parts of the reference speech signal (301) with the highest intensity summed over their length are matched with corresponding signal parts of the test speech signal (302), then the matching continues with signal parts of the reference speech signal (301) with decreasing intensity summed over their length.

2. The method according to claim 1, wherein the reference speech signal (301) and the test speech signal (302) are each pre-filtered by a bandpass filter, in particular by a bandpass filter with a frequency range that corresponds to the audio band.

3. The method according to claim 1 or 2,
wherein a performance measure is computed for each pair of matched signal parts (303, 304), the performance measure being in particular the maximum of the cross-correlation of the matched signal parts (303, 304) normalized by the signal powers of the matched signal parts (303, 304).

4. The method according to claim 3, wherein a pair of matched signal parts is deemed to have insufficient performance if its performance measure lies beneath a pre-set threshold value.

5. The method according to claim 3, wherein a pair of matched signal parts is deemed to have insufficient performance if its performance measure is significantly lower than the performance measures of other pairs of matched signal parts (303, 304).

6. The method according to claim 4 or 5,
wherein each signal part of a pair of matched signal parts with deemed insufficient performance is rematched.

7. The method according to one of the preceding claims, wherein the reference speech signal (101) has a reference speech spectrum (X) and the test speech signal (102) has a test speech spectrum (Y) and wherein the method further comprises the following steps:

    - identifying a number of perceptually dominant frequency sub-bands (bi) in one of the reference speech spectrum (X) and the test speech spectrum (Y),
    - computing an intensity scaling factor (c_i) for each identified sub-band (bi) by minimizing a measure of the intensity difference between those parts of the reference speech spectrum (X) and of the test speech spectrum (Y) that correspond to the respective sub-band (b_i).
    - multiplying the test speech spectrum (Y) with each intensity scaling factor (c_i) thereby generating a number of scaled test speech spectra (Y_i),
    - selecting one scaled test speech spectrum, and
    - computing the difference between the selected scaled test speech spectrum (Y_{sel}) and the reference speech spectrum (X), the difference being indicative of the speech quality of the test speech signal (102).

8. The method according to claim 7, wherein the measure of the intensity difference is given by the squared intensity difference or the global maximum of the intensity difference between those parts of the reference speech spectrum (X) and of the test speech spectrum (Y) that correspond to the respective sub-band (b_i).

9. The method according to claim 7 or 8,
wherein the number of perceptually dominant sub-bands (bi) of one of the reference speech spectrum (X) and the test signal spectrum (Y) is identified by determining the local maxima in a perceptual repre-

sentation of the respective spectrum and by selecting a predetermined range of frequencies around each local maximum.

10. The method according to claim 9, wherein the predetermined range of frequencies is determined by the local minima bordering the respective local maximum with the predetermined range of frequencies being in particular smaller or equal to 4 Bark.

11. The method according to claim 9 or 10, wherein the perceptual representation of the respective spectrum is obtained by transforming the respective spectrum to a loudness spectrum.

12. The method according to one of the claims 7 to 11, wherein a measure of the intensity difference between the reference speech spectrum (X) and a respective scaled test speech spectrum (Y_i) is computed for each scaled test speech spectrum (Y_i) and wherein the scaled test speech spectrum is selected that yields the lowest measure of the intensity difference.

13. The method according to claim 12, wherein the measure of the intensity difference is given by the squared intensity difference or the global maximum of the intensity difference between the reference speech spectrum (X) and a respective scaled test speech spectrum (Y_i).

14. A method according to one of the preceding claims, wherein the method further comprises the following steps:

- for at least one missed or interrupted signal part in the test speech signal (402) computing the signal intensities of the signal parts adjacent to the missed or interrupted signal part,
- deriving an expected signal intensity for the at least one missed or interrupted signal part from the computed signal intensities of the adjacent signal parts of the test speech signal (402),
- computing a measure of the perceived distortion by comparing the actual intensity of the at least one missed or interrupted signal part in the test speech signal (402) with the derived expected intensity for the at least one missed or interrupted signal part,
- computing a measure of the actual distortion by comparing the reference speech signal (401) with the test speech signal (402), and
- combining the measure of the perceived distortion with the measure of the actual distortion to generate a combined measure of distortion indicative of the speech quality of the test speech signal (402).

15. The method according to claim 14, wherein the expected signal intensity of the at least one missed or interrupted signal part is derived from the computed signal intensities of the adjacent signal parts of the test speech signal (402) by means of interpolation, in particular by means of linear interpolation and/or spline interpolation.

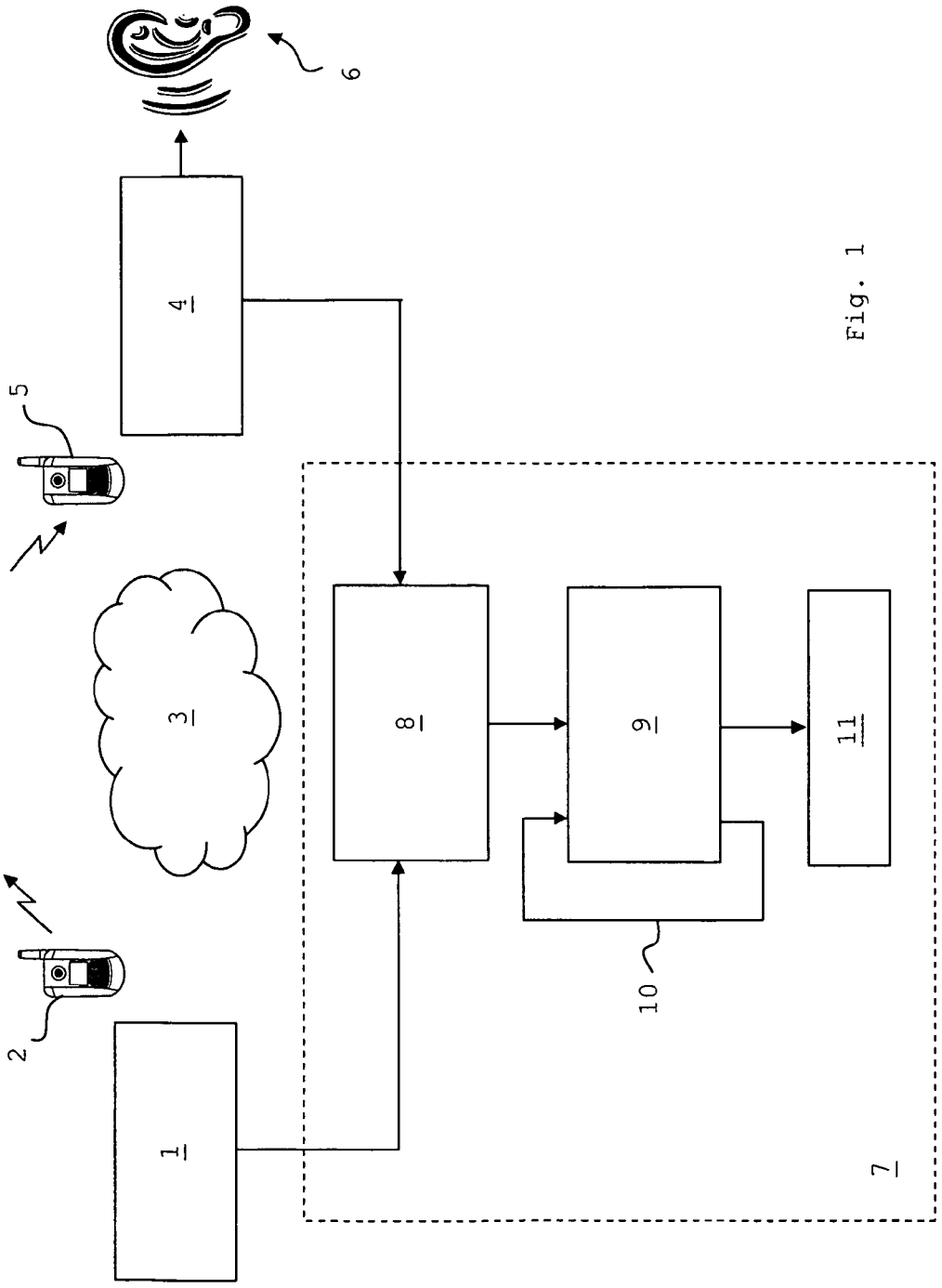

Fig. 1

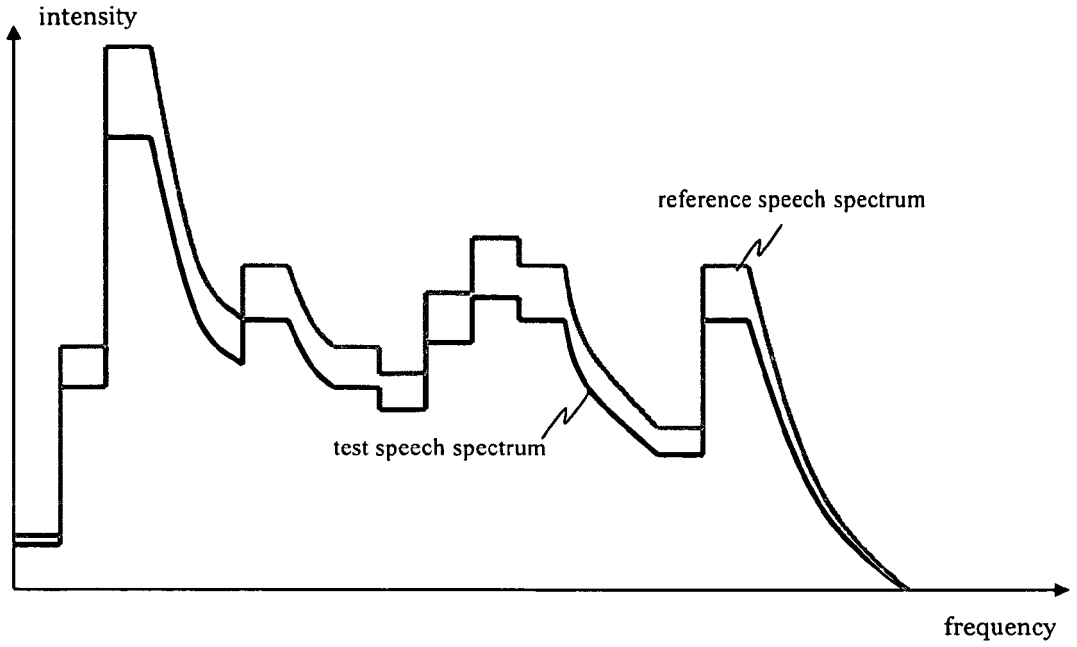
intensity
reference speech spectrum
test speech spectrum
frequency

Fig. 2

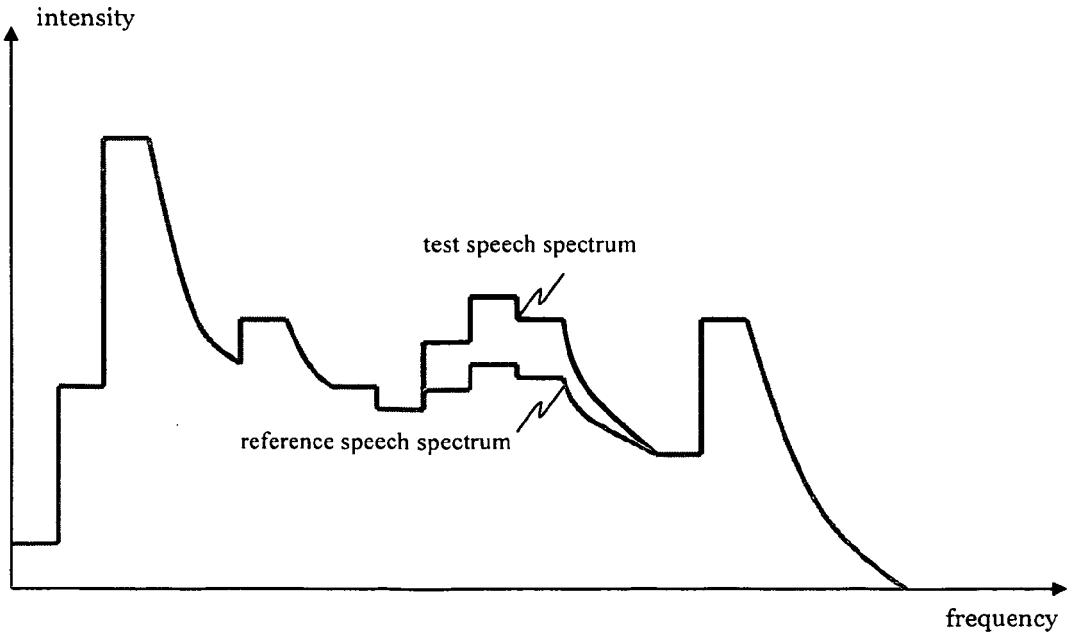
intensity
test speech spectrum
reference speech spectrum
frequency

Fig. 3

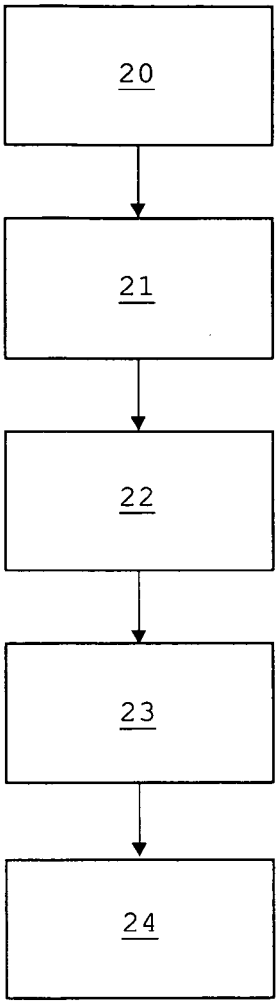

Fig. 4

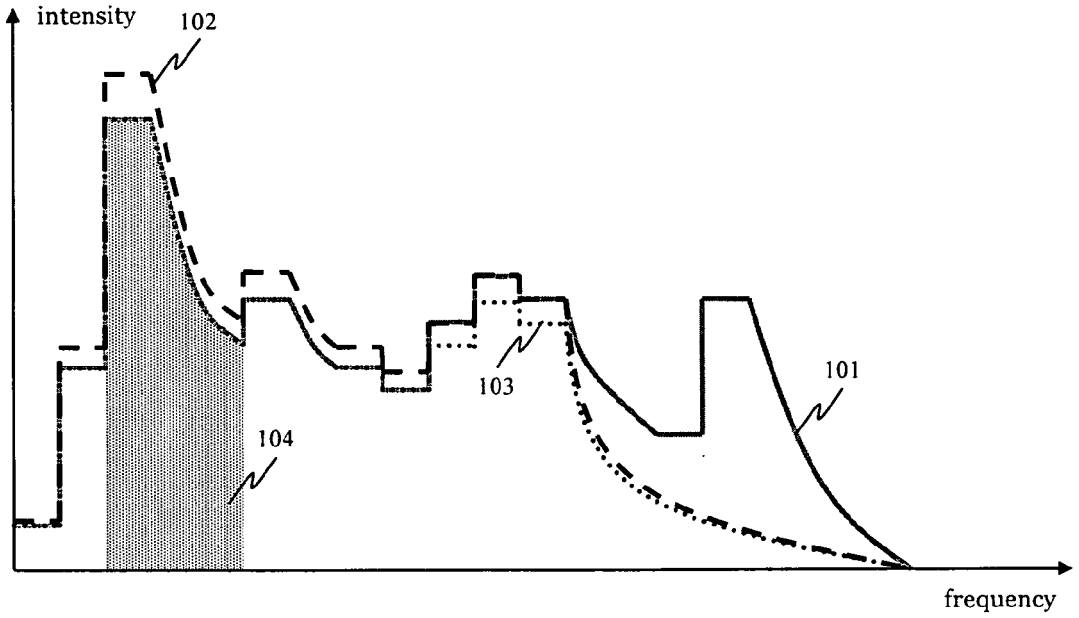

Fig. 5

Stand der Technik

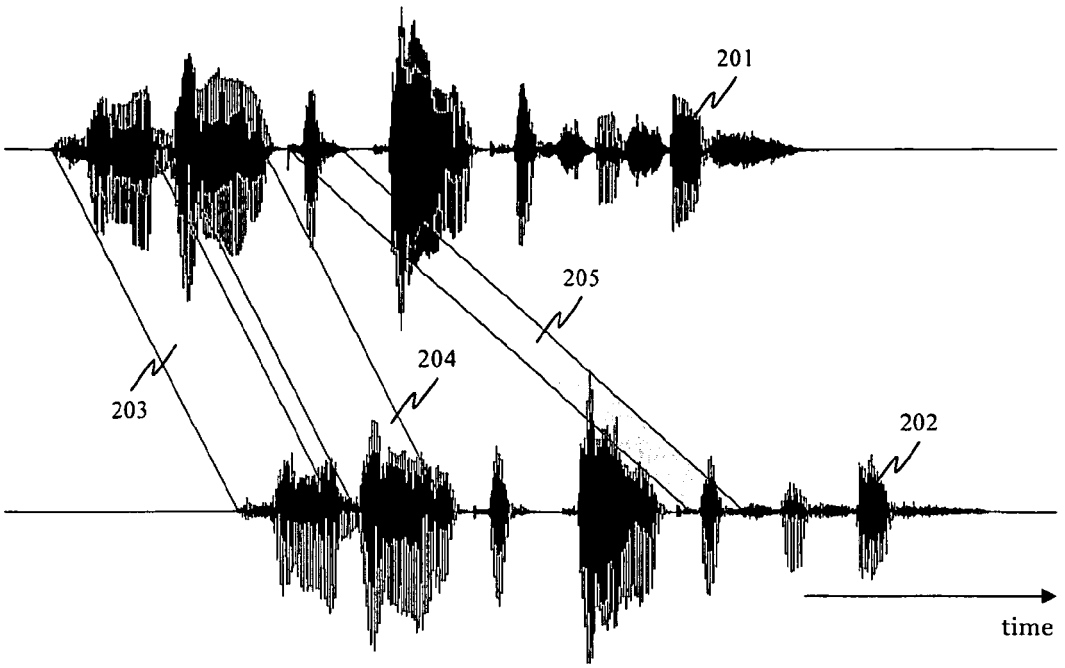

Fig. 6

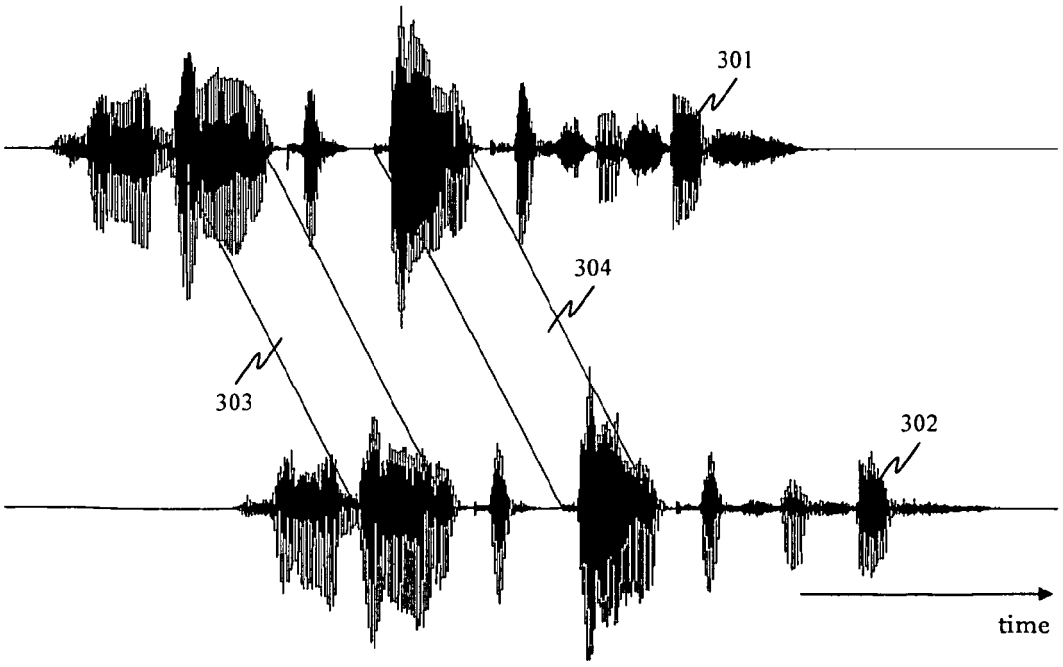

Fig. 7

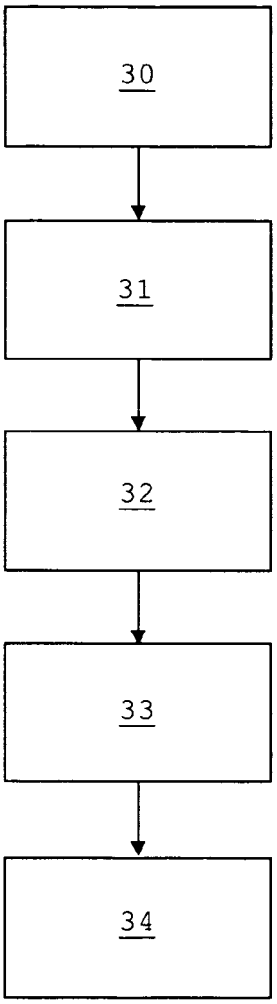

Fig. 8

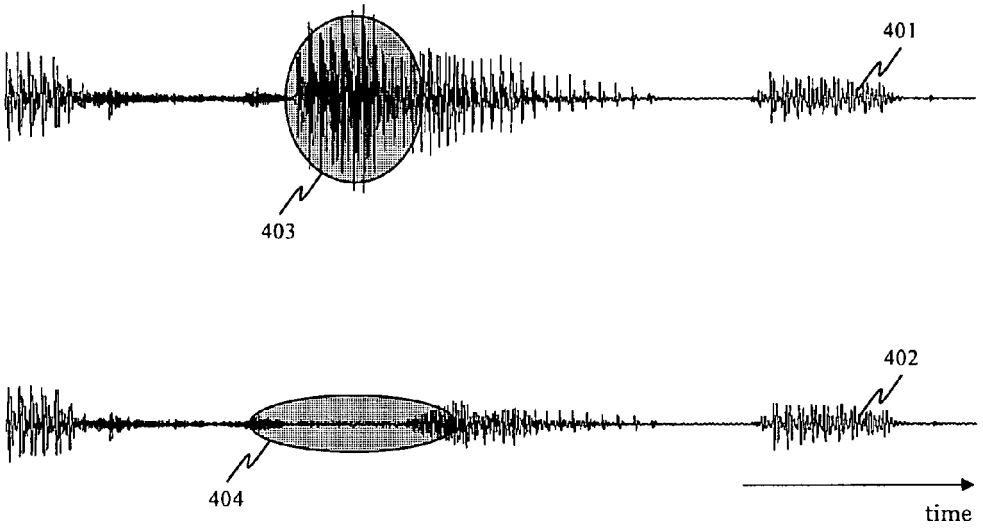

Fig. 9

Europäisches Patentamt
European Patent Office
Office européen des brevets

EUROPEAN SEARCH REPORT

Application Number
EP 12 00 0483

DOCUMENTS CONSIDERED TO BE RELEVANT

Category	Citation of document with indication, where appropriate, of relevant passages	Relevant to claim	CLASSIFICATION OF THE APPLICATION (IPC)
A	EP 1 104 924 A1 (KONINKL KPN NV [NL]) 6 June 2001 (2001-06-06) * abstract; figures 1, 2a-c, 3a-c, 6 * * paragraph [0064] - paragraph [0079] *	1-15	INV. G10L19/00
A	EP 0 946 015 A1 (ASCOM INFRASYS AG [CH] ASCOM SCHWEIZ AG [CH]) 29 September 1999 (1999-09-29) * abstract; figure 2 *	1-15	
A	DE 198 40 548 A1 (DEUTSCHE TELEKOM AG [DE]) 2 March 2000 (2000-03-02) * abstract; figure 2b *	1-15	
			TECHNICAL FIELDS SEARCHED (IPC)
			G10L

The present search report has been drawn up for all claims

Place of search	Date of completion of the search	Examiner
Munich	3 May 2012	Zimmermann, Elko

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

1

EPO FORM 1503 03.82 (P04C01)

ANNEX TO THE EUROPEAN SEARCH REPORT ON EUROPEAN PATENT APPLICATION NO. EP 12 00 0483

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

03-05-2012

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
EP 1104924 A1	06-06-2001	AU 1145801 A	12-06-2001
		EP 1104924 A1	06-06-2001
		EP 1240644 A1	18-09-2002
		US 7139705 B1	21-11-2006
		WO 0141127 A1	07-06-2001
EP 0946015 A1	29-09-1999	AT 376295 T	15-11-2007
		CA 2326138 A1	07-10-1999
		EP 0946015 A1	29-09-1999
		ES 2296327 T3	16-04-2008
		TW 428380 B	01-04-2001
		US 7212815 B1	01-05-2007
		WO 9950991 A1	07-10-1999
DE 19840548 A1	02-03-2000	AT 253765 T	15-11-2003
		CA 2305652 A1	09-03-2000
		DE 19840548 A1	02-03-2000
		EP 1048025 A1	02-11-2000
		US 7013266 B1	14-03-2006
		WO 0013173 A1	09-03-2000

EPO FORM P0459

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

REFERENCES CITED IN THE DESCRIPTION

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

Non-patent literature cited in the description

- **PAULUS, E. ; ZWICKER, E.** Programme zur automatischen Bestimmung der Lautheit aus Terzpegeln oder Frequenzgruppenpegeln. *Acustica,* 1972, vol. 27 (5 **[0005] [0022]**
- **BEERENDS, J.G. ; HEKSTRA, A.P. ; RIX, A.W. ; HOLLIER, M.P.** Perceptual Evaluation of Speech Quality (PESQ), the new ITU standard for end-to-end speech quality assessment. Part I - Time alignment. *J. Audio. Eng. Soc.,* October 2002, vol. 50 (10 **[0007]**
- **BEERENDS, J.G. ; HEKSTRA, A.P. ; RIX, A.W. ; HOLLIER, M.P.** Peceptial 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 (10 **[0007]**
- **BEERENDS, J.G. ; STEMERDINK, J.A.** A perceptual speech quality measure based on a psychoacoustic sound representation. *J. Audio Eng. Soc.,* December 1994, vol. 42 (3 **[0007]**
- TOSQA - Telecommunication objective speech quality assessment. *COM12-34-E, Geneva,* December 1997 **[0007]**
- Methods for subjective determination of transmission quality. *ITU-T Recommendation P.800,* 1996 **[0008]**
- Objective measurement of active speech level. *ITU-T Recommendation P.56,* 1993 **[0046]**