(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
1 December 2005 (01.12.2005)

PCT

(10) International Publication Number
WO 2005/114656 A1

(51) International Patent Classification[7]: G10L 21/02

(21) International Application Number:
PCT/EP2004/050816

(22) International Filing Date: 14 May 2004 (14.05.2004)

(25) Filing Language: English

(26) Publication Language: English

(71) Applicant *(for all designated States except US)*: **LOQUENDO S.p.A.** [IT/IT]; Via Valdellatorre, 4, I-10149 TORINO (IT).

(72) Inventors; and
(75) Inventors/Applicants *(for US only)*: **GEMELLO, Roberto** [IT/IT]; Via Valdellatorre, 4, I-10149 Torino (IT). **MANA, Franco** [IT/IT]; Via Valdellatorre, 4, I-10149 Torino (IT).

(74) Agents: **JORIO, Paolo** et al.; STUDIO TORTA S.r.l., Via Viotti, 9, I-10121 TORINO (IT).

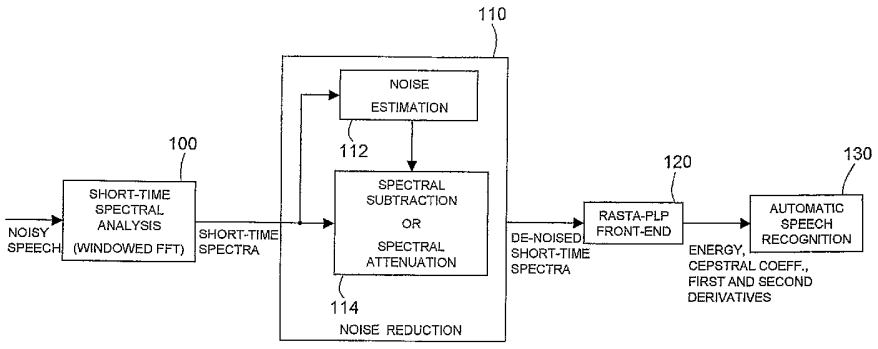
(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BW, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NA, NI, NO, NZ, OM, PG, PH, PL, PT, RO, RU, SC, SD, SE, SG, SK, SL, SY, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, YU, ZA, ZM, ZW.

(84) Designated States *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HU, IE, IT, LU, MC, NL, PL, PT, RO, SE, SI, SK, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:
— *with international search report*

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

(54) **Title:** NOISE REDUCTION FOR AUTOMATIC SPEECH RECOGNITION



(57) **Abstract:** Disclosed herein is a noise reduction method for automatic speech recognitionl. A noise reduction method for automatic speech recognition, including: computing a magnitude spectrum ($Y_k(m)$ of a noisy speech containing a clean speech to be recognized and noise affecting the clean speech; computing a power spectrum ($|Y_k(m|^2$) of the noisy speech; computing an estimate ($|X_k(m)|^2$ of a power spectrum of the clean speech; computing an estimate ($D_k(m|^2$) of a power spectrum of the noise; computing an estimate ($\xi k(m)^2$) of an a priori signal-to-noise ratio as a function of the estimate ($|X_k,(m)|^2$) of the power spectrum of the clean speech and the estimate ($|D_k(m)|^2$of the power spectrum of the noise; computing an estimate ($\gamma_k(m)$) of an a posteriori signal-to-noise ratio as a function of the power spectrum ($|Y_k(m)|^2$) of the noisy speech and the estimate ($|D_k(m)|^2$) of the power spectrum of the noise; computing an attenuation gain ($G_k(m)$) as a function of the estimate ($\xi k(m)$) of the a priori signal-to-noise ratio and the estimate ($\gamma_k(m)$) of the a posteriori signal-to-noise ratio; and computing an estimate ($|X_k(m)$) of a magnitude spectrum of the clean speech as a function of the magnitude spectrum ($|Y_k(m)$) of the noisy speech and the attenuation gain ($G_k(m)$). Computing the estimates ($\xi_k(m),\gamma_k(m)$) of the a priori and the a posteriori signal-to-noise ratios includes computing a noise weighting factor ($\alpha(m)$) for weighting the estimate ($D_k(m)^2$) of the power spectrum of the noise in the computation of the estimates ($\xi_k(m),\gamma_k(m)$) of the a priori and the a posteriori signal-to-noise ratios; computing a spectral flooring factor ($\beta(m)$) for flooring the estimates ($\xi_k(m)$, $\gamma_k(m)$)of the a priori and the a posteriori signal-to-noise ratios; and computing the estimates ($\xi k(m)$, $\gamma_k(m)$) of the a priori and the a posteriori signal-to-noise ratios also as a function of the noise weighting factor ($\alpha(m)$ and the spectral flooring factor ($\beta(m)$).

WO 2005/114656 A1

NOISE REDUCTION FOR AUTOMATIC SPEECH RECOGNITION

TECHNICAL FIELD OF THE INVENTION

The present invention relates in general to noise reduction for automatic speech recognition, and in particular to a noise reduction method and system based on Spectral Attenuation Technique, and to an automatic speech recognition system.

BACKGROUND ART

Figure 1 shows a block diagram of common sources of speech degradation. As may be appreciated, speech from the desired speaker (block 10) is degraded by environmental noise, namely voices of other speakers nearby (block 20) and background noise (block 30), and by communication channel noise and distortion (blocks 40 and 50). Noise reduction techniques (block 60) for automatic speech recognition (block 70) can reduce the (nearly stationary) background noise and the channel noise, whereas non-stationary noise and interfering voices are much more difficult to be eliminated.

Figure 2 shows a block diagram of an automatic speech recognition system. As may be appreciated, the noisy speech to be recognized is inputted to a short-time spectral analysis (windowed FFT) block 100 which outputs short-time spectra which is in turn inputted to a noise reduction block 110. The de-noised short-time spectra are inputted to a RASTA-PLP Front-End 120, which outputs the total energy of the speech signal, the cepstral coefficients, and the first and second derivatives of the total energy and of the cepstral coefficients, which are all inputted to an automatic speech recognition block 130.

RASTA-PLP Front-End 120 implements a technique known as "RelAtive SpecTrAl Technique", which is an

improvement of the traditional PLP (Perceptual Linear Prediction) method and consists in a special filtering of the different frequency channels of a PLP analyzer. The previous filtering is done to make speech analysis less sensitive to the slowly changing or steady-state factors in speech. The RASTA method replaces the conventional critical-band short-term spectrum in PLP and introduces a less sensitive spectral estimate. For a more detailed description of a RASTA processing, reference may be made to H. Hermansky and N. Morgan, *RASTA Processing of Speech, IEEE Transactions on Speech and Audio Processing*, Vol. 2 No. 4, October 1994.

The noise reduction block 110 performs an environmental noise estimate 112 based on the short-time spectra and then an environmental noise reduction 114 based on the short-time spectra and the estimated noise, by using either a so-called "Spectral Subtraction Technique" or a so-called "Spectral Attenuation Technique".

The aforementioned techniques will be described in detail hereinafter by denoting the power spectrum of the noisy speech by $|Y_k(m)|^2$, the power spectrum of the clean speech by $|X_k(m)|^2$, the power spectrum of the additive noise by $|D_k(m)|^2$, and the estimate of a quantity by symbol "^", and wherein k indexes the spectral lines of the spectra and m indexes the time windows within which the noisy speech is processed for noise reduction.

Spectral Subtraction Technique is described in N. Virag, *Single Channel Speech Enhancement Based on Masking Properties of the Human Auditory System*, IEEE Transaction on Speech and Audio Processing, Vol. 7, No. 2, March 1999, which deals with the problem of noise reduction for speech recognition and discloses the use of a noise overestimation or oversubtraction factor and a spectral flooring factor.

In particular, the Spectral Subtraction Technique is based on the principle of reducing the noise by subtracting an estimate $|\hat{D}_k(m)|^2$ of the power spectrum of the additive noise from the power spectrum $|Y_k(m)|^2$ of the noisy speech, thus obtaining an estimate $|\hat{X}_k(m)|^2$ of the power spectrum of the clean speech:

$$|\hat{X}_k(m)|^2 = \begin{cases} |Y_k(m)|^2 - \alpha(m)|\hat{D}_k(m)|^2 & \text{if } |Y_k(m)|^2 - \alpha(m)|\hat{D}_k(m)|^2 > \beta(m)|Y_k(m)|^2 \\ \beta(m)|Y_k(m)|^2 & \text{otherwise} \end{cases} \quad (1)$$

wherein $\alpha(m)$ is the noise overestimation factor, $\beta(m)$ is the spectral flooring factor.

In particular, the residual noise spectrum consists of peaks and valleys with random occurrences, and the overestimation factor $\alpha(m)$ and the spectral flooring factor $\beta(m)$ have been introduced to reduce the spectral excursions.

In detail, the overestimation factor $\alpha(m)$ has been introduced to "overestimate" the noise spectrum, i.e., in other words the overestimation factor $\alpha(m)$ subtracts an overestimation of the noise over the whole spectrum, whereas the spectral flooring factor $\beta(m)$ prevents the spectral lines of the estimate $|\hat{X}_k(m)|^2$ of the power spectrum of the clean speech from descending below a lower bound ($\beta(m)|Y_k(m)|^2$), thereby "filling-in" the deep valleys surrounding narrow peaks (from the enhanced spectrum). In fact, occasional negative estimates of the enhanced power spectrum can occur and in such cases, the negative spectral lines are floored to zero or to some minimal value (floor). Reducing the spectral excursions of noise peaks as compared to when the negative components are set to zero, reduces the amount of musical noise. Essentially by reinserting the broadband noise (noise floor), the remnants of the noise peaks are

"masked" by neighboring components of comparable magnitude.

A variant of this technique in known as "Wiener Spectral Subtraction Technique", which is similar to the previous one but is derived from the optimal filter theory. The estimate $|\hat{X}_k(m)|^2$ of the power spectrum of the clean speech is the following:

$$|\hat{X}_k(m)|^2 = \begin{cases} \dfrac{\left[|Y_k(m)|^2 - \alpha(m)|\hat{D}_k(m)|^2\right]^2}{|Y_k(m)|^2} & \text{if } |Y_k(m)|^2 - \alpha(m)|\hat{D}_k(m)|^2 > \beta(m)|Y_k(m)|^2 \\ \beta(m)|Y_k(m)|^2 & \text{otherwise} \end{cases} \quad (2)$$

An improvement to the Spectral Subtraction Techniques is disclosed in V. Schless, F. Class, *SNR-Dependent flooring and noise overestimation for joint application of spectral subtraction and model combination*, ICSLP 1998, which proposes to make the noise overestimation factor $\alpha(m)$ and the spectral flooring factor $\beta(m)$ functions of the global signal-to-noise ratio $SNR(m)$.

Spectral Attenuation Technique, instead, is based on the principle of suppressing the noise by applying a suppression rule, or a non-negative real-valued gain G_k, to each spectral line k of the magnitude spectrum $|Y_k(m)|$ of the noisy speech, in order to compute an estimate $|\hat{X}_k(m)|$ of the magnitude spectrum of the clean speech according to the following formula: $|\hat{X}_k(m)| = G_k(m)|Y_k(m)|$.

Many suppression rules have been proposed, and probably one of the most important rules is the so-called Ephraim-Malah spectral attenuation log rule, which is described in Y. Ephraim and D. Malah, *Speech Enhancement Using a Minimum Min-Square Error Log-Spectral Amplitude Estimator*, IEEE Transaction on

Acoustics, Speech, and Signal Processing, Vol. ASSP-33, No. 2, pp. 443-445, 1985.

Ephraim-Malah gain $G_k(m)$ is defined as:

$$G_k(m) = \frac{\xi_k(m)}{1+\xi_k(m)} exp\left(\frac{1}{2}\int_{\nu_k}^{\infty}\frac{e^{-t}}{t}dt\right) \quad (3)$$

where:

- $\xi_k(m)$ is a so-called *a priori* signal-to-noise ratio relating to the *k*-th spectral line and is defined as follows:

$$\xi_k(m) = \frac{|X_k(m)|^2}{|D_k(m)|^2} \quad (4)$$

- $\nu_k(m)$ is defined as:

$$\nu_k(m) = \frac{\xi_k(m)}{1+\xi_k(m)}\gamma_k(m) \quad (5)$$

- $\gamma_k(m)$ is a so-called *a posteriori* signal-to-noise ratio relating to the *k*-th spectral line and is defined as follows:

$$\gamma_k(m) = \frac{|Y_k(m)|^2}{|D_k(m)|^2} \quad (6)$$

Computation of the *a posteriori* signal-to-noise ratio $\gamma_k(m)$ requires the knowledge of the power spectrum $|D_k(m)|^2$ of the additive noise, which is not available. An estimate $|\hat{D}_k(m)|^2$ of the power spectrum of the additive noise can be obtained with a noise estimate as described in H. G. Hirsch, C. Ehrlicher, *Noise Estimation Techniques for Robust Speech Recognition*, ICASSP 1995,

pp.153-156.

Thus, an estimate $\hat{\gamma}_k(m)$ of the *a posteriori* signal-to-noise ratio may be computed as follows:

$$\hat{\gamma}_k(m) = \frac{|Y_k(m)|^2}{|\hat{D}_k(m)|^2} \qquad (7)$$

Computation of the *a priori* signal-to-noise ratio $\xi_k(m)$ requires the knowledge of the power spectrum $|X_k(m)|^2$ of the clean speech, which is not available. An estimate $\hat{\xi}_k(m)$ of the *a priori* signal-to-noise ratio can be computed by using a decision-directed approach as described in Y. Ephraim and D. Malah, *Speech Enhancement Using a Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator*, IEEE Trans. Acoustic, Speech, and Signal Processing, Vol. ASSP-32, Mo. 6, pp. 1109-1121, 1984, and as follows:

$$\hat{\xi}_k(m) = \eta(m)\frac{|\hat{X}_k(m-1)|^2}{|\hat{D}_k(m-1)|^2} + [1-\eta(m)]\max[0, \hat{\gamma}_k(m)-1], \quad \eta(m) \in [0,1) \qquad (8)$$

where $\eta(m)$ is a weighting coefficient for appropriately weighting the two terms in the formula.

The Ephraim-Malah gain $G_k(m)$ may then be computed as a function of the estimate $\hat{\xi}_k(m)$ of the *a priori* signal-to-noise ratio and of the estimate $\hat{\gamma}_k(m)$ of the *a posteriori* signal-to-noise ratio according to formula (3).

An application of the Spectral Attenuation Technique is disclosed in US-A-2002/0002455, which relates to a speech enhancement system receiving noisy speech characterized by a spectral amplitude spanning a plurality of frequency bins and producing enhanced speech by modifying the spectral amplitude of the noisy

speech without affecting the phase thereof. In particular, the speech enhancement system includes a core estimator that applies to the noisy speech one of a first set of gains for each frequency bin; a noise adaptation module that segments the noisy speech into noise-only and signal-containing frames, maintains a current estimate of the noise spectrum and an estimate of the probability of signal absence in each frequency bin; and a signal-to-noise ratio estimator that measures *a posteriori* signal-to-noise ratio and estimates *a priori* signal-to-noise ratio based on the noise estimate. Each one of the first set of gains is based on *a priori* signal-to-noise ratio, as well as the probability of signal absence in each bin and a level of aggression of the speech enhancement. A soft decision module computes a second set of gains that is based on *a posteriori* signal-to-noise ratio and *a priori* signal-to-noise ratio, and the probability of signal absence in each frequency bin.

Another application of the Spectral Attenuation Techniques is disclosed in WO-A-01/52242, which relates to a multi-band spectral subtraction scheme which can be applied to a variety of speech communication systems, such as hearing aids, public address systems, teleconference systems, voice control systems, or speaker phones, and which comprises a multi band filter architecture, noise and signal power detection, and gain function for noise reduction. The gain function for noise reduction consists of a gain scale function and a maximum attenuation function providing a predetermined amount of gain as a function of signal-to-noise ratio and noise. The gain scale function is a three-segment piecewise linear function, and the three piecewise linear sections of the gain scale function include a first section providing maximum expansion up to a first

knee point for maximum noise reduction, a second section providing less expansion up to a second knee point for less noise reduction, and a third section providing minimum or no expansion for input signals with high signal-to-noise ratio to minimize distortion. The maximum attenuation function can either be a constant or equal to the estimated noise envelope. When used in hearing aid applications, the noise reduction gain function is combined with the hearing loss compensation gain function inherent to hearing aid processing.

Automatic speech recognition performed by using the known noise reduction methods described above is affected by some technical problems which prevents it from being really effective. In particular, Spectral Subtraction Technique and Wiener Spectral Subtraction Technique are affected by the so-called "musical noise", which is introduced in the power spectrum $|X_k(m)|^2$ of the clean speech by the aforementioned flooring, according to which negative values are set to a flooring value $\beta(m)|Y_k(m)|^2$ in order to avoid occurrence of negative subtraction results. In particular, the flooring introduces discontinuities in the spectrum that are perceived as annoying musical noises and degrade the performances of an automatic speech recognition system.

Spectral Attenuation Technique implementing the Ephraim-Malah attenuation rule is a very good technique for the so-called speech enhancement, i.e., noise reduction for a human listener, but it introduces some spectral distortion on voice parts that are acceptable for humans but very critical for an automatic speech recognition system.

OBJECT AND SUMMARY OF THE INVENTION

The aim of the present invention is therefore to provide a noise reduction method for automatic speech

recognition and which, at the same time, reduces the musical noise in the power spectrum of the clean speech.

This aim is achieved by the present invention in that it relates to a noise reduction method for automatic speech recognition, as defined in claim 1, to an automatic speech recognition system, as defined in claim 12, and to a computer program product, as defined in claim 13.

The present invention meets the aforementioned needs as it makes use of a Spectral Attenuation Technique instead of a Spectral Subtraction Technique, thus removing the problem of musical noise, and the use of a modified Ephraim-Malah spectral attenuation rule implemented reduces the spectral distortion introduced by the original rule in the voice parts of the signals, thus obtaining better performances when used in an automatic speech recognition system.

BRIEF DESCRIPTION OF THE DRAWINGS

For a better understanding of the present invention, a preferred embodiment, which is intended purely by way of example and is not to be construed as limiting, will now be described with reference to the attached drawings, wherein:

- Figure 1 shows a block diagram of common sources of speech degradation;
- Figure 2 shows a block diagram of noise reduction for automatic speech recognition;
- Figures 3 and 4 show plots of a noise overestimation factor and a spectral flooring factor as a function of a global signal-to-noise ratio and used in the noise reduction method according to the present invention;
- Figure 5 shows a standard Ephraim-Malah spectral attenuation rule; and

- Figures 6-10 show a modified Ephraim-Malah spectral attenuation rule according to the present invention at different global signal-to-noise ratio.

DETAILED DESCRIPTION OF PREFERRED EMBODIMENTS OF THE INVENTION

The following discussion is presented to enable a person skilled in the art to make and use the invention. Various modifications to the embodiments will be readily apparent to those skilled in the art, and the generic principles herein may be applied to other embodiments and applications without departing from the spirit and scope of the present invention. Thus, the present invention is not intended to be limited to the embodiments shown, but is to be accorded the widest scope consistent with the principles and features disclosed herein and defined in the attached claims.

The present invention relates to an automatic speech recognition system including a noise reduction system based on the Spectral Attenuation Technique, and in particular on the Ephraim-Malah spectral attenuation rule, wherein the global formula of the gain $G_k(\gamma_k, \xi_k)$ is unchanged, whereas the estimates of the *a priori* and the *a posteriori* signal-to-noise ratios $\hat{\xi}_k(m)$, $\hat{\gamma}_k(m)$ are modified by making them dependent on a noise weighting factor $\alpha(m)$ and on a spectral flooring factor $\beta(m)$, as follows:

$$\hat{\gamma}_k(m) = max\left(\frac{|Y_k(m)|^2}{\alpha(m)|\hat{D}_k(m)|^2} - 1, \beta(m) \right) + 1 \qquad (9)$$

$$\hat{\xi}_k(m) = \max\left(\eta(m) \frac{|\hat{X}_k(m-1)|^2}{\alpha(m)|\hat{D}_k(m-1)|^2} + (1-\eta(m))[\hat{\gamma}_k(m)-1], \beta(m) \right), \quad \eta(m) \in [0,1) \qquad (10)$$

where:

- $|Y_k(m)|^2$ is the k-th spectral line of the power spectrum of the noisy speech;
- $|\hat{X}_k(m)|^2$ is the k-th spectral line of the estimate of the power spectrum of the clean speech;
- $|\hat{D}_k(m)|^2$ is the k-th spectral line of the estimate of the power spectrum of the additive noise;
- $\hat{\xi}_k(m)$ is the estimate of the *a priori* signal-to-noise ratio relating to the k-th spectral line;
- $\hat{\gamma}_k(m)$ is the estimate of the *a posteriori* signal-to-noise ratio relating to the k-th spectral line;
- $\alpha(m)$ is the noise weighting factor for weighting, namely overestimating or underestimating, the estimate $|\hat{D}_k(m)|^2$ of the power spectrum of the noise in the computation of the estimates $\hat{\xi}_k(m)$, $\hat{\gamma}_k(m)$ of the *a priori* and the *a posteriori* signal-to-noise ratios;;
- $\beta(m)$ is the spectral flooring factor for flooring the estimates $\hat{\xi}_k(m)$, $\hat{\gamma}_k(m)$ of the *a priori* and the *a posteriori* signal-to-noise ratios; and
- $\eta(m)$ is a weighting coefficient for appropriately weighting the two terms in formula (10).

The noise weighting factor $\alpha(m)$ and the spectral flooring factor $\beta(m)$ are a function of the global signal-to-noise ratio $SNR(m)$, which is defined as:

$$SNR(m) = 10\log_{10}\left(\frac{\sum_k |Y_k(m)|^2}{\sum_k |\hat{D}_k(m)|^2}\right) \qquad (11)$$

Figures 3 and 4 show a preferred development of the noise weighting factor $\alpha(m)$ and the spectral flooring

factor $\beta(m)$ versus the global signal-to-noise ratio $SNR(m)$. The noise weighting factor $\alpha(m)$ and the spectral flooring factor $\beta(m)$ are piece-wise linear functions and may be mathematically defined as follows:

$$\alpha(m)=\begin{cases}1.5 & if \quad SNR(m)<0\\ 1.5-\dfrac{(1.5-0.001)}{20}\cdot SNR(m) & if \quad 0\le SNR(m)\le 20\\ 0.001 & if \quad SNR(m)>20\end{cases} \qquad (12)$$

$$\beta(m)=\begin{cases}0.01 & if \quad SNR(m)<0\\ \dfrac{(1.0-0.01)}{20}\cdot SNR(m) & if \quad 0\le SNR(m)\le 20\\ 1.0 & if \quad SNR(m)>20\end{cases} \qquad (13)$$

The values indicated in formulas (12) and (13) are intended purely by way of example and are not to be construed as limiting. In general, other values could be usefully employed, while maintaining the general development of the noise weighting factor $\alpha(m)$ and of the spectral flooring factor $\beta(m)$ versus the global signal-to-noise ratio $SNR(m)$.

In particular, the noise weighting factor $\alpha(m)$ versus the global signal-to-noise ratio $SNR(m)$ should have a first substantially constant value when the global signal-to-noise ratio $SNR(m)$ is lower than a first threshold, a second substantially constant value lower than the first substantially constant value when the global signal-to-noise ratio $SNR(m)$ is higher than a second threshold, and values decreasing from the first substantially constant value to the second substantially constant value when the global signal-to-noise ratio $SNR(m)$ increases from the first threshold to the second threshold.

The spectral flooring factor $\beta(m)$ versus the global signal-to-noise ratio $SNR(m)$ should have a first substantially constant value when the global signal-to-

noise ratio $SNR(m)$ is lower than a first threshold, a second substantially constant value higher than the first substantially constant value when the global signal-to-noise ratio $SNR(m)$ is higher than a second threshold, and values increasing from the first substantially constant value to the second substantially constant value when the global signal-to-noise ratio $SNR(m)$ increases from the first threshold to the second threshold. The developments may be piece-wise lines, as shown in Figures 3 and 4, or may be continuous curved lines similar to those in Figures 3 and 4, i.e., curved lines wherein the intermediate non-constant stretch is linear, as in Figures 3 and 4, or curved, e.g., a cosine-like or a sine-like curve, and transitions from the intermediate non-constant stretch to the constant stretches is rounded or smoothed.

The estimate $\left|\hat{D}_k(m)\right|^2$ of the power spectrum of the noisy speech in formulas (9), (10) and (11) is computed by means of a first-order recursion as disclosed in the aforementioned *Noise Estimation Techniques for Robust Speech Recognition.*

Preferably, the first-order recursion may be implemented in conjunction with a standard energy-based Voice Activity Detector, which is well-known system which detects presence or absence of speech based on a comparison of the total energy of the speech signal with an adaptive threshold and outputs a Boolean flag (VAD) having a "true" value when voice is present and a "false" value when voice is absent. When a standard energy-based Voice Activity Detector is used, the estimate $\left|\hat{D}_k(m)\right|^2$ of the power spectrum of the noisy speech may be computed as follows:

$$\left|\hat{D}_k(m)\right|^2 = \begin{cases} \lambda\left|\hat{D}_k(m-1)\right|^2 + (1-\lambda)\left|Y_k(m)\right|^2 & \text{if } \left\{\left|\left|Y_k(m)\right|^2 - \left|\hat{D}_k(m)\right|^2\right| \le \mu\sigma(m)\right\} \wedge \{VAD = false\} \\ \left|\hat{D}_k(m-1)\right|^2 & \text{otherwise} \end{cases} \tag{14}$$

where λ is a weighting factor which controls the update speed of the recursion and ranges between 0 and 1, preferably has a value of 0.9, μ is a multiplication factor which controls the allowed dynamics of the noise and preferably has a value of 4.0, and $\sigma(m)$ is the noise standard deviation, estimated as follows:

$$\sigma^2(m) = \lambda\sigma^2(m-1) + (1-\lambda)\left(\left|Y_k(m)\right|^2 - \left|\hat{D}_k(m)\right|^2\right)^2 \tag{15}$$

Figure 5 shows the standard Ephraim-Malah spectral attenuation rule (G_k, $\xi_k(m)$ and $\gamma_k(m)$ computed according to formulas (3), (7) and (8)), whereas Figures 6-10 show the modified Ephraim-Malah spectral attenuation rule according to the present invention (G_k, $\xi_k(m)$ and $\gamma_k(m)$ computed according to formulas (3), (10) and (9)) at different global signal-to-noise ratios $SNR(m)$ (0, 5, 10, 15 and 20 dB). It may be appreciated by the skilled person that the effect of the introduced modification is a gradual reduction of the attenuation produced by the original gain in areas where the *a posteriori* $\gamma_k(m)$ signal-to-noise ratios is high, as the global signal-to-noise ratios $SNR(m)$ increases.

A large experimental work has been performed to validate the invention, and some results, which may be useful to highlight the features of the invention, are hereinafter reported.

In particular, experiments were conducted with an automatic speech recognition system, using noise reduction with the standard Ephraim-Malah spectral attenuation and with the noise reduction proposed in the

invention. The automatic speech recognition system has been trained for the target languages using large, domain and task independent corpora, not collected in noisy environments and without added noise.

The experiment was performed on the Aurora3 corpus, that is a standard corpus defined by the ETSI Aurora Project for noise reduction tests, and made of connected digits recorded in car in several languages (Italian, Spanish and German). An high mismatch test set and a noisy component of the training set (used as test set) were employed.

The modification of the Ephraim-Malah spectral attenuation rule according to the invention produces an average error reduction of 28.9% with respect to the state of the art Wiener Spectral Subtraction, and an average error reduction of 22.9% with respect to the standard Ephraim-Malah Spectral Attenuation Rule. The average error reduction with respect to no de-noising is 50.2%.

Finally, it is clear that numerous modifications and variants can be made to the present invention, all falling within the scope of the invention, as defined in the appended claims.

CLAIMS

1. A noise reduction method for automatic speech recognition, including:

- computing a magnitude spectrum ($|Y_k(m)|$) of a noisy speech containing a clean speech to be recognized and noise affecting the clean speech;

- computing a power spectrum ($|Y_k(m)|^2$) of the noisy speech;

- computing an estimate ($|\hat{X}_k(m)|^2$) of a power spectrum of the clean speech;

- computing an estimate ($|\hat{D}_k(m)|^2$) of a power spectrum of the noise;

- computing an estimate ($\hat{\xi}_k(m)$) of an *a priori* signal-to-noise ratio as a function of the estimate ($|\hat{X}_k(m)|^2$) of the power spectrum of the clean speech and the estimate ($|\hat{D}_k(m)|^2$) of the power spectrum of the noise;

- computing an estimate ($\hat{\gamma}_k(m)$) of an *a posteriori* signal-to-noise ratio as a function of the power spectrum ($|Y_k(m)|^2$) of the noisy speech and the estimate ($|\hat{D}_k(m)|^2$) of the power spectrum of the noise;

- computing an attenuation gain ($G_k(m)$) as a function of the estimate ($\hat{\xi}_k(m)$) of the *a priori* signal-to-noise ratio and the estimate ($\hat{\gamma}_k(m)$) of the *a posteriori* signal-to-noise ratio;

- computing an estimate ($|\hat{X}_k(m)|$) of a magnitude spectrum of the clean speech as a function of the magnitude spectrum ($|Y_k(m)|$) of the noisy speech and the attenuation gain ($G_k(m)$);

characterized in that computing the estimates ($\hat{\xi}_k(m), \hat{\gamma}_k(m)$) of the *a priori* and the *a posteriori* signal-to-noise ratios includes:

- computing a noise weighting factor ($\alpha(m)$) for weighting the estimate ($|\hat{D}_k(m)|^2$) of the power spectrum of the noise in the computation of the estimates ($\hat{\xi}_k(m)$,

$\hat{\gamma}_k(m)$) of the *a priori* and the *a posteriori* signal-to-noise ratios;

- computing a spectral flooring factor ($\beta(m)$) for flooring the estimates ($\hat{\xi}_k(m)$, $\hat{\gamma}_k(m)$) of the *a priori* and the *a posteriori* signal-to-noise ratios; and

- computing the estimates ($\hat{\xi}_k(m)$, $\hat{\gamma}_k(m)$) of the *a priori* and the *a posteriori* signal-to-noise ratios also as a function of the noise weighting factor ($\alpha(m)$) and the spectral flooring factor ($\beta(m)$).

2. A noise reduction method as claimed in claim 1, wherein the noise weighting factor ($\alpha(m)$) and the spectral flooring factor ($\beta(m)$) are computed as a function of a global signal-to-noise ratio ($SNR(m)$).

3. A noise reduction method as claimed in claim 2, wherein the noise weighting factor ($\alpha(m)$) versus the global signal-to-noise ratio ($SNR(m)$) has a first substantially constant value when the global signal-to-noise ratio ($SNR(m)$) is lower than a first threshold, a second substantially constant value lower than the first substantially constant value when the global signal-to-noise ratio ($SNR(m)$) is higher than a second threshold, and decreasing values when the global signal-to-noise ratio ($SNR(m)$) ranges between the first and the second thresholds.

4. A noise reduction method as claimed in claim 3, wherein the noise weighting factor ($\alpha(m)$) decreases linearly when the global signal-to-noise ratio ($SNR(m)$) ranges between the first and the second thresholds.

5. A noise reduction method as claimed in any one of the claims 2 to 4, wherein the spectral flooring factor ($\beta(m)$) versus the global signal-to-noise ratio ($SNR(m)$) has a first substantially constant value when the global signal-to-noise ratio ($SNR(m)$) is lower than a first threshold, a second substantially constant value higher than the first substantially constant value when

the global signal-to-noise ratio ($SNR(m)$) is higher than a second threshold, and increasing values when the global signal-to-noise ratio ($SNR(m)$) ranges between the first and the second thresholds.

6. A noise reduction method as claimed in claim 5, wherein the spectral flooring factor ($\beta(m)$) increases linearly when the global signal-to-noise ratio ($SNR(m)$) ranges between the first and the second thresholds.

7. A noise reduction method as claimed in any one of the preceding claims, wherein the estimate ($\hat{\gamma}_k(m)$) of the *a posteriori* signal-to-noise ratio is computed as follows:

$$\hat{\gamma}_k(m) = \max\left(\frac{|Y_k(m)|^2}{\alpha(m)|\hat{D}_k(m)|^2} - 1, \beta(m)\right) + 1$$

wherein:

- $\hat{\gamma}_k(m)$ is the estimate of the *a posteriori* signal-to-noise ratio for the k-th spectral line;
- $|Y_k(m)|^2$ is the k-th spectral line of the power spectrum of the noisy speech;
- $|\hat{D}_k(m)|^2$ is the k-th spectral line of the estimate of the power spectrum of the noise;
- $\alpha(m)$ is the noise weighting factor;
- $\beta(m)$ is the spectral flooring factor;
- k indexes spectral lines of the spectra; and
- m indexes time windows within which the noisy speech is processed for noise reduction.

8. A noise reduction method as claimed in any one of the preceding claims, wherein the estimate ($\hat{\xi}_k(m)$) of the *a priori* signal-to-noise ratio is computed as follows:

$$\hat{\xi}_k(m)=\max\left\{\eta\frac{\left|\hat{X}_k(m-1)\right|^2}{\alpha(m)\left|\hat{D}_k(m-1)\right|^2}+(1-\eta)[\hat{\gamma}_k(m)-1],\beta(m)\right\},\ \eta(m)\in[0,1)$$

wherein:

- $\hat{\xi}_k(m)$ is the estimate of the *a priori* signal-to-noise ratio for the is k-th spectral line;
- $\hat{\gamma}_k(m)$ is the estimate of the *a posteriori* signal-to-noise ratio for the k-th spectral line of the power spectrum of the noisy speech;
- $\left|\hat{X}_k(m)\right|^2$ is the k-th spectral line of the estimate of the power spectrum of the clean speech;
- $\left|\hat{D}_k(m)\right|^2$ is the k-th spectral line of the estimate of the power spectrum of the noise;
- $\alpha(m)$ is the noise weighting factor;
- $\beta(m)$ is the spectral flooring factor;
- k indexes spectral lines of the spectra; and
- m indexes time windows within which the noisy speech is processed for noise reduction.

9. A noise reduction method as claimed in any one of the preceding claims, wherein the attenuation gain ($G_k(m)$) is computed as follows:

$$G_k(m)=\frac{\hat{\xi}_k(m)}{1+\hat{\xi}_k(m)}\exp\left(\frac{1}{2}\int_{\nu_k}^{\infty}\frac{e^{-t}}{t}dt\right)$$

wherein:

- $G_k(m)$ is the Ephraim-Malah gain for the k-th spectral line;
- $\hat{\xi}_k(m)$ is the estimate of the *a priori* signal-to-noise ratio for the k-th spectral line;
- $\hat{\gamma}_k(m)$ is the estimate of the *a posteriori* signal-to-noise ratio for the k-th spectral line;

- $v_k(m) = \frac{\hat{\xi}_k(m)}{1+\hat{\xi}_k(m)}\hat{\gamma}_k(m)$
- k indexes spectral lines of the spectra; and
- m indexes time windows within which the noisy speech is processed for noise reduction.

10. A noise reduction method as claimed in any one of the preceding claims, wherein the estimate ($|\hat{D}_k(m)|^2$) of the power spectrum of the noise is computed as follows:

$$|\hat{D}_k(m)|^2 = \begin{cases} \lambda|\hat{D}_k(m-1)|^2 + (1-\lambda)|Y_k(m)|^2 & \text{if } \left\{\left||Y_k(m)|^2 - |\hat{D}_k(m)|^2\right| \le \mu\sigma(m)\right\} \wedge \{VAD = false\} \\ |\hat{D}_k(m-1)|^2 & \text{otherwise} \end{cases}$$

where:

- $|\hat{D}_k(m)|^2$ is the k-th spectral line of the estimate of the power spectrum of the noise;
- $|Y_k(m)|^2$ is the k-th spectral line of the power spectrum of the noisy speech;
- λ is a weighting factor which controls the update speed of the recursion,
- μ is a multiplication factor which controls the allowed dynamics of the noise, and
- $\sigma(m)$ is the noise standard deviation, estimated as follows:

$$\sigma^2(m) = \lambda\sigma^2(m-1) + (1-\lambda)\left(|Y_k(m)|^2 - |\hat{D}_k(m)|^2\right)^2$$

11. A noise reduction method as claimed in claim 2, wherein the global signal-to-noise ratio ($SNR(m)$) is computed as follows:

$$SNR(m) = 10\log_{10}\left(\frac{\sum_k |Y_k(m)|^2}{\sum_k |\hat{D}_k(m)|^2}\right)$$

where:

- $SNR(m)$ is the global signal-to-noise ratio;
- $|\hat{D}_k(m)|^2$ is the k-th spectral line of the estimate of the power spectrum of the noise; and
- $|Y_k(m)|^2$ is the k-th spectral line of the power spectrum of the noisy speech.

12. An automatic speech recognition system including a noise reduction system configured to implement the method according to any one of the foregoing claims.

13. A computer program product comprising a computer program code able, when loaded in a processing system, to implement the method according to any one of claims 1 to 11.

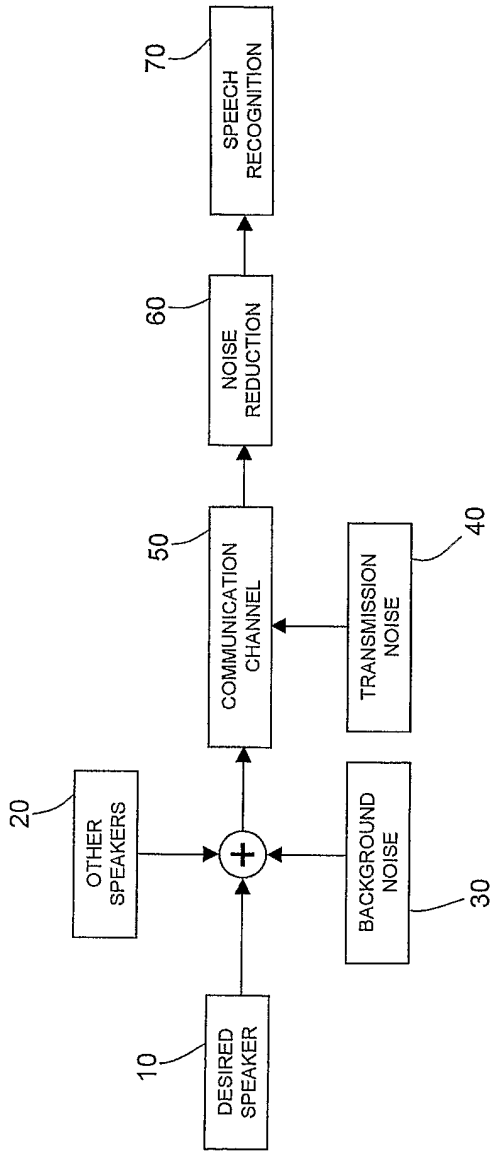

70
SPEECH RECOGNITION
60
NOISE REDUCTION
50
COMMUNICATION CHANNEL
40
TRANSMISSION NOISE
20
OTHER SPEAKERS
30
BACKGROUND NOISE
10
DESIRED SPEAKER

Fig. 1

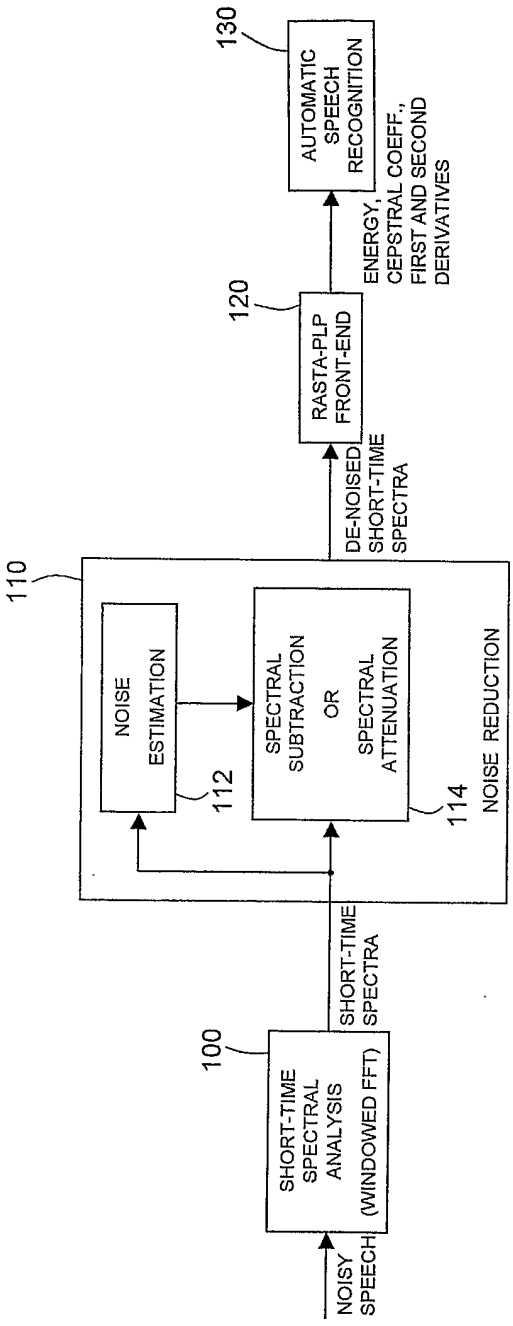

NOISY SPEECH
SHORT-TIME SPECTRAL ANALYSIS (WINDOWED FFT)
100
SHORT-TIME SPECTRA
110
NOISE ESTIMATION
112
SPECTRAL SUBTRACTION OR SPECTRAL ATTENUATION
114
NOISE REDUCTION
DE-NOISED SHORT-TIME SPECTRA
120
RASTA-PLP FRONT-END
ENERGY, CEPSTRAL COEFF., FIRST AND SECOND DERIVATIVES
130
AUTOMATIC SPEECH RECOGNITION

Fig. 2

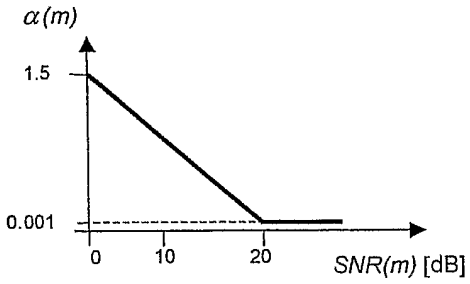


Fig.3

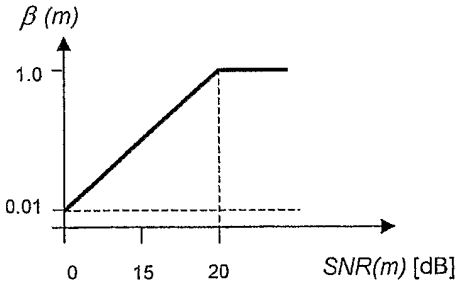


Fig.4

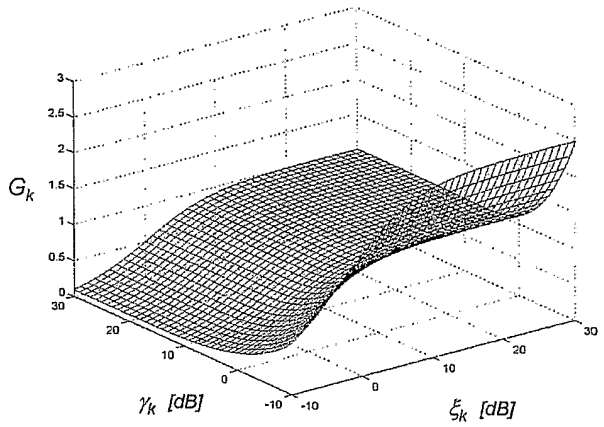

G_k
3
2.5
2
1.5
1
0.5
0
30
20
10
0
-10
γ_k [dB]
-10
0
10
20
30
ξ_k [dB]

Fig.5

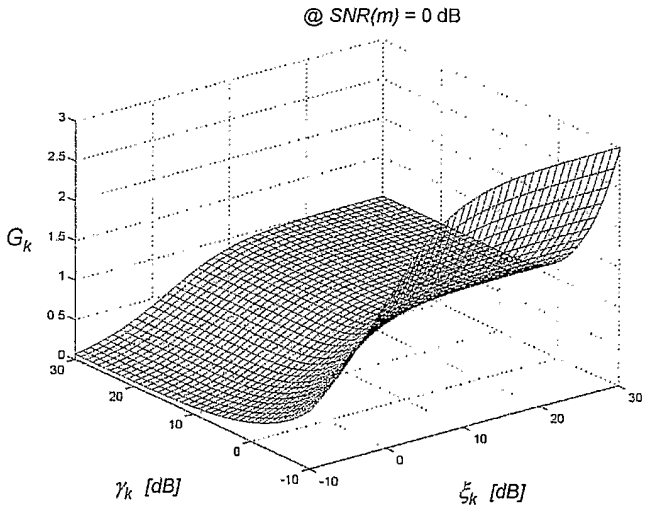

@ SNR(m) = 0 dB
G_k
3
2.5
2
1.5
1
0.5
0
30
20
10
0
-10
γ_k [dB]
-10
0
10
20
30
ξ_k [dB]

Fig.6

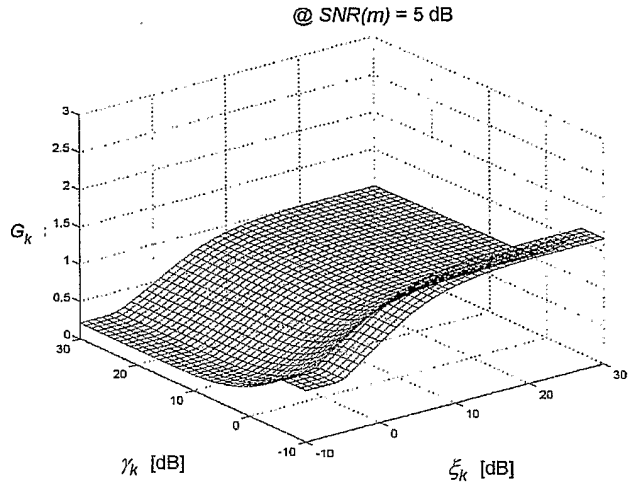

@ SNR(m) = 5 dB
G_k
3
2.5
2
1.5
1
0.5
0
30
20
10
0
-10
γ_k [dB]
-10
0
10
20
30
ξ_k [dB]

Fig.7

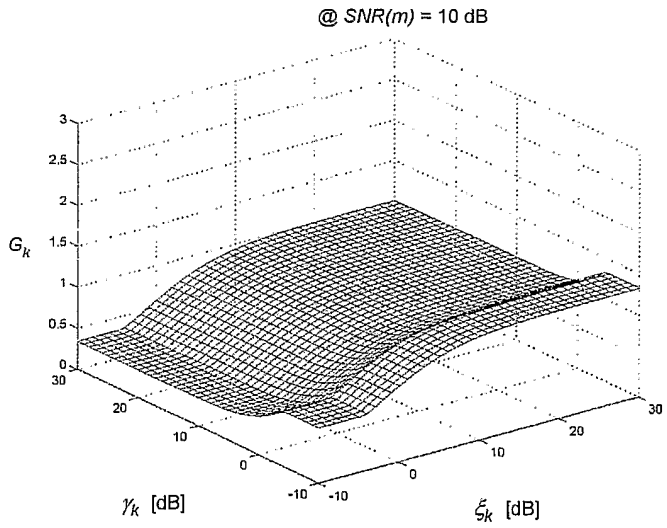

@ SNR(m) = 10 dB
G_k
3
2.5
2
1.5
1
0.5
0
30
20
10
0
-10
γ_k [dB]
-10
0
10
20
30
ξ_k [dB]

Fig.8

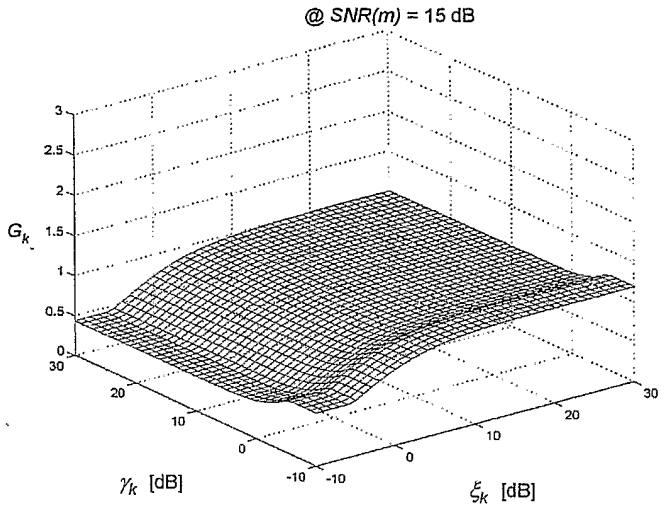

@ SNR(m) = 15 dB
G_k
3
2.5
2
1.5
1
0.5
0
30
20
10
0
-10
γ_k [dB]
-10
0
10
20
30
ξ_k [dB]

Fig.9

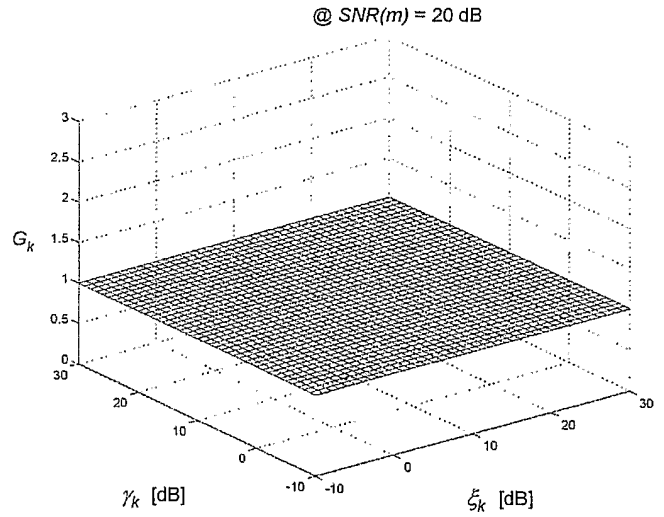

@ SNR(m) = 20 dB
G_k
3
2.5
2
1.5
1
0.5
0
30
20
10
0
-10
γ_k [dB]
-10
0
10
20
30
ξ_k [dB]

Fig.10

INTERNATIONAL SEARCH REPORT

International Application No
[redacted]/EP2004/050816

A. CLASSIFICATION OF SUBJECT MATTER
IPC 7 G10L21/02

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
IPC 7 G10L

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practical, search terms used)
EPO-Internal

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category °	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	KATO M ET AL: "A wideband noise suppressor for the amr wideband speech codec" SPEECH CODING, 2002, IEEE WORKSHOP PROCEEDINGS, 6 October 2002 (2002-10-06), - 9 October 2002 (2002-10-09) pages 156-158, XP010647245 chapter 2 -----	1-13
A	MARTIN R ET AL: "Optimized estimation of spectral parameters for the coding of noisy speech" INTERNATIONAL CONFERENCE ON ACOUSTICS, SPEECH AND SIGNAL PROCESSING, ICASSP 2000, vol. 3, 5 June 2000 (2000-06-05), pages 1479-1482, XP010507630 chapter 3 ----- -/--	1-13

[X] Further documents are listed in the continuation of box C.

[ ] Patent family members are listed in annex.

° Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier document but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art.

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
4 November 2004	03/12/2004

Name and mailing address of the ISA	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL – 2280 HV Rijswijk Tel. (+31–70) 340–2040, Tx. 31 651 epo nl, Fax: (+31–70) 340–3016	Bourdier, R

INTERNATIONAL SEARCH REPORT

International Application No
PCT/EP2004/050816

C.(Continuation) DOCUMENTS CONSIDERED TO BE RELEVANT

Category °	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	EPHRAIM Y ET AL: "SPEECH ENHANCEMENT USING A MINIMUM MEAN-SQUARE ERROR LOG-SPECTRAL AMPLITUDE ESTIMATOR" IEEE TRANSACTIONS ON ACOUSTICS, SPEECH AND SIGNAL PROCESSING, IEEE INC. NEW YORK, US, vol. ASSP-33, no. 2, April 1985 (1985-04), pages 443-445, XP000931203 ISSN: 0096-3518 cited in the application the whole document	1-13
A	SCHWAB M ET AL INSTITUTE OF ELECTRICAL AND ELECTRONICS ENGINEERS: "Robust noise estimation applied to different speech estimators" CONFERENCE RECORD OF THE 37TH. ASILOMAR CONFERENCE ON SIGNALS, SYSTEMS, & COMPUTERS. PACIFIC GROOVE, CA, NOV. 9 - 12, 2003, ASILOMAR CONFERENCE ON SIGNALS, SYSTEMS AND COMPUTERS, NEW YORK, NY : IEEE, US, vol. VOL. 1 OF 2. CONF. 37, 9 November 2003 (2003-11-09), pages 1904-1907, XP010701031 ISBN: 0-7803-8104-1 chapter 2.2	1-13
A	MEDINA C A ET AL INSTITUTE OF ELECTRICAL AND ELECTRONICS ENGINEERS: "Robust speaker verification in colored noise environment" CONFERENCE RECORD OF THE 37TH. ASILOMAR CONFERENCE ON SIGNALS, SYSTEMS, & COMPUTERS. PACIFIC GROOVE, CA, NOV. 9 - 12, 2003, ASILOMAR CONFERENCE ON SIGNALS, SYSTEMS AND COMPUTERS, NEW YORK, NY : IEEE, US, vol. VOL. 1 OF 2. CONF. 37, 9 November 2003 (2003-11-09), pages 1890-1893, XP010702165 ISBN: 0-7803-8104-1 Chapter II.B	1-13
A	VOLKER SCHLESS FRITZ CLASS: "SNR-DEPENDENT FLOORING AND NOISE OVERESTIMATION FOR JOINT APPLICATION OF SPECTRAL SUBTRACTION AND MODEL COMBINATION -" ICSLP 1998, October 1998 (1998-10), page P138, XP007000096 cited in the application Chapters 2 and 4	1-13