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

(11) EP 1 944 754 A1

(12) EUROPEAN PATENT APPLICATION

(43) Date of publication:
16.07.2008 Bulletin 2008/29

(51) Int Cl.:
G10L 11/04 $^{(2006.01)}$

(21) Application number: **07000568.1**

(22) Date of filing: **12.01.2007**

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

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

(72) Inventors:
- **Krini, Mohamed**
  89073 - Ulm (DE)
- **Schmidt, Gerhard**
  89081 - Ulm (DE)

(74) Representative: **Bertsch, Florian Oliver et al**
Kraus & Weisert
Patent- und Rechtsanwälte
Thomas-Wimmer-Ring 15
80539 München (DE)

Remarks:
Amended claims in accordance with Rule 137(2) EPC.

(54) Speech fundamental frequency estimator and method for estimating a speech fundamental frequency

(57) The present invention relates to a speech fundamental frequency estimator (1100) which is configured for receiving a first set of values ($\tilde{Y}_1$) and a second set of values ($\tilde{Y}_2$), the first set of values ($\tilde{Y}_1$) being a frequency domain representation of a first set of time domain signal values (y_1) within a first time interval (t_1) and the second set of values ($\tilde{Y}_2$) being a frequency domain representation of a second set of time domain signal values (y_2) within a second time interval (t_2), the second time interval (t_2) being later than and offset from the first time interval (t_1). Furthermore, the speech fundamental frequency estimator (1100) comprises a first power density spectrum calculator (1102) which is configured for storing a version of the first set of values ($\tilde{Y}_1$) and being configured for providing values of a first power density spectrum

$$(\hat{S}_{\tilde{y}\tilde{y}_d}(\Omega_\mu, n))$$

by multiplying the stored version of the first set of values ($\tilde{Y}_1$) with a conjugate complex version of the second set of values ($\tilde{Y}_2$). In addition the speech fundamental estimator (1100) comprises a second power density spectrum calculator (1104) being configured for providing values of a second power density spectrum

$$(\hat{S}_{\tilde{y}\tilde{y}}(\Omega_\mu, n))$$

by multiplying a version of the second set of values ($\tilde{Y}_2$) with a complex conjugate verisin of the second set of values ($\tilde{Y}_2$). Finally, the speech fundamental frequency estimator (1100) includes an analyzer 1(106) which is configured for determining the speech fundamental frequency estimate (fp(n)) on the basis of the values of the first power density spectrum

$$(\hat{S}_{\tilde{y}\tilde{y}_d}(\Omega_\mu, n))$$

and the values of the second power density spectrum

$$(\hat{S}_{\tilde{y}\tilde{y}}(\Omega_\mu, n)).$$

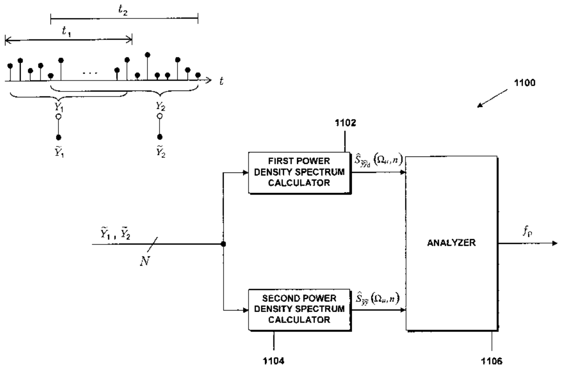

t_2
t_1
t
Y_1
Y_2
$\tilde{Y}_1$
$\tilde{Y}_2$
1100
1102
FIRST POWER DENSITY SPECTRUM CALCULATOR
$\hat{S}_{\tilde{y}\tilde{y}_d}(\Omega_\mu, n)$
$\tilde{Y}_1$, $\tilde{Y}_2$
N
ANALYZER
f_p
SECOND POWER DENSITY SPECTRUM CALCULATOR
$\hat{S}_{\tilde{y}\tilde{y}}(\Omega_\mu, n)$
1104
1106

FIG. 11A

Printed by Jouve, 75001 PARIS (FR)

Description

[0001] This invention relates to speech analysis systems and especially to a speech fundamental frequency estimator and a method for estimating a speech fundamental frequency.

Background of the invention

[0002] An estimation of the speech fundamental frequency is necessary in various applications:

- For example, partial speech recognition (model-based noise suppression) can be accomplished by the estimated speech fundamental frequency of distorted speech signals in order to obtain an improvement of speech quality.

- In order to name a further example a rough preselection of model parameters can be performed by a speech recognizer in speech recognition systems using the temporal average of this frequency. Thus, the recognition rate of a speech recognizer can be increased significantly.

[0003] For further fields of application reference is made to the following literature:

- K. Fellbaum: Sprachverarbeitung und Sprachübertragung, Springer, Berlin, Deutschland, 1984
- D.K. Freeman, G. Cosier, C.B. Southcott und I. Boyd: The Voice Activity Detector for the PAN-European Digital Cellular Mobile Telephone Service, Proceed. of the Intern. Conf. on Acoust., Speech, and Signal Process., Vol. 1, pages 369-372, 1989
- W. Hess: Pitch Determination of Speech Signals, Springer, Berlin, Deutschland, 1983
- P. Vary, R. Martin: Digital Speech Transmission, John Wiley & Sons, Chichester, England, 2006
- P. Vary, U. Heute, W. Hess: Digitale Sprachsignalverarbeitung, Teubner, Stuttgart, Deutschland, 1998

[0004] Numerous methods for estimation of the speech fundamental frequency exist. A group of methods which is based on a DFT-transform (DFT = discrete Fourier transform) of the input signal is of special importance. Such methods can be integrated in hands-free speech assistance systems with a multi-rate signal processing in a low-cost way as the DFT-transform is already calculated for other algorithms, as, for example, a noise reduction or an echo compensation.

[0005] In order to describe the relevant state of the art in more detail, a typical multi-rate system is described which can be used, for example, in order to perform a speech signal improvement (noise reduction, speech reconstruction). Following, several further fields of application are presented in which an estimation of the reliable speech fundamental frequency is of importance.

[0006] In voiced speech portions the corresponding spectrum shows distinct amplitude peaks which are located equidistantly in frequency (see for example Fig. 1). The distance between two amplitude peaks represents herein the speech fundamental frequency which is dependent of the speaker. With men this frequency varies between 80 Hz and 150 Hz, women and children, in contrast, have a higher speech fundamental frequency which varies between 150 Hz and 300 Hz with women, respectively between 200 Hz and 600 Hz with children. A good, sure and reliable estimation of the speech fundamental frequency is often not easy to obtain. Mainly difficulties in detecting low speech fundamental frequencies arise wherein especially men have in most cases a low speech fundamental frequency.

[0007] In Figure 2 a block diagram of a multi-rate system for speech reconstruction with an analysis and a synthesis filter bank for the signal processing is shown. The speech fundamental frequency estimation is shown as a separate functional block. The aim of such an application is to extract parameters from a distorted speech signal y(n) as, for example, the spectral envelope, the type of stimulation (voiced/unvoiced) and the speech fundamental frequency $f_p(n)$. Subsequently an undistorted speech signal x(n) is resynthesized from these parameters. For this purpose a very precise and reliable estimation of the speech fundamental frequency is necessary. The output signal x(n) after the synthesis filter bank should be nearly without error, the following condition is therefore very desirable:

$$x(n) \approx s(n), \tag{1}$$

s(n) denotes herein the undisturbed speech signal.

[0008] A sure estimation of the speech fundamental frequency is also of great importance in speech recognition systems. Figure 3 shows a block diagram of a signal analysis system with subsequent feature extraction and speech fundamental frequency estimation, in order to perform a speech recognition. An adequate estimation of the speech fundamental frequency can, for example, contribute to significantly improve the recognition rates of the speech recognizer.

[0009] Basically there is a broad variety of application fields in which a reliable estimation of the speech fundamental frequency is necessary. However a detailed description of such applications would go beyond the scope of this description. Thus, reference is made to the following literature:

- E. Hänsler, G. Schmidt: Acoustic Echo and Noise Control - A Practical Approach, John Wiley & Sons, Hoboken, New Jersey, USA, 2004
- E. Hänsler, G. Schmidt: Topics in Acoustic Echo and Noise Control - Selected Methods for the Cancellation of Acoustic Echoes, the Reduction of Background Noise, and Speech Processing, Springer, Berlin, Deutschland, 2006
- P. Vary, R. Martin: Digital Speech Transmission, John Wiley & Sons, Chichester, England, 2006
- P. Vary, U. Heute, W. Hess: Digitale Sprachsignalverarbeitung, Teubner, Stuttgart, Deutschland, 1998

[0010] In literature a broad variety of different algorithms for a determination of the speech fundamental frequency estimation exist as for example:

- **Analysis in the cepstral domain** - In the case speech generation is modelled as a source-filter-model (see also J. Deller, J. Hansen, J. Proakis: Discrete-Time Processing of Speech Signals, IEEE-Press, New York, USA, 1993) voiced sounds can be described as a convolution of a periodic stimulation signal with the impulse response of the vocal tract. In the spectral domain the convolution becomes the product of the Fourier transforms of both portions. If the Logarithm is taken the product becomes an addition of the separate components. After a further transform (inverse Fourier transform) the cepstral domain is reached. In this domain it is possible to distinguish the spectrally comparatively slowly varying frequency response of the vocal tract from the fundamental frequency of the stimulation signal. Further details can be found for example in W. Hess: Pitch Determination of Speech Signals, Springer, Berlin, Deutschland, 1983.

- **Harmonic Product-Spectrum** - Another method to estimate the speech fundamental frequency is the so-called Harmonic Product-Spectrum. Herein the product over several equidistant sampling points of the absolute value of the spectrum is calculated. The product becomes maximal in the case the increment (via frequency) corresponds to just the speech fundamental frequency (respectively a multiple thereof). Further details can be found for example in M. R. Schroeder: Period Histogram and Product Spectrum: New Methods for Fundamental Frequency Measurements, J. Acoust. Soc. Am., Vol. 43, Nr. 4, pages 829-834, 1968.

- **Analysis of the short-time autocorrelation -** In voiced speech passages the first side lobe of the short-time autocorrelation with an offset just corresponds to the speech fundamental period.

[0011] Other methods as the ones mentioned above also exist. The description of each single algorithm would, however, be far beyond the possibilities given the present description. Therefore reference is made to further literature as, for example, K. Fellbaum: Sprachverarbeitung und Sprachübertragung, Springer, Berlin, Deutschland, 1984 or D.K. Freeman, G. Cosier, C.B. Southcott and I. Boyd: The Voice Activity Detector for the PAN-European Digital Cellular Mobile Telephone Service, Proceed. of the Intern. Conf. on Acoust., Speech, and Signal Process., Vol. 1, Seiten 369-372, 1989 or W. Hess: Pitch Determination of Speech Signals, Springer, Berlin, Deutschland, 1983. The approach mentioned last in the above listing has become very popular as it provides the advantage that already determined short-time DFT-portions of an input signal, which are calculated for other applications, can be further used and thus a numerical effort can be reduced.

[0012] However, the above mentioned approach of the state of the art also has significant disadvantages. Especially the orders of the DFT (i.e. the DFT-block length) used for other purposes are often to little as to provide a reliable estimation of the speech fundamental frequency for low voices.

[0013] Accordingly, a need exists to provide a speech fundamental frequency estimator and a method for estimating a speech fundamental frequency which allow a more precise estimation of the speech fundamental frequency.

[0014] This need is met by the features of the independent claims. In the dependent claims further embodiment of the inventions are described.

[0015] According to a first aspect of the invention the speech fundamental frequency estimator is configured for receiving a first set of values and a second set of values, the first set of values being a frequency domain representation of a first set of time domain signal values within a first time interval and the second set of values being a frequency domain representation of a second set of time domain signal values within a second time interval, the second time interval being later than and offset from the first time interval, the speech fundamental frequency estimator comprising:

- a first power density spectrum calculator being configured for storing a version of the first set of values and being configured for providing values of a first power density spectrum by multiplying the stored version of the first set of values with a complex conjugate version of the second set of values;

- a second power density spectrum calculator being configured for providing values of a second power density spectrum by multiplying a version of the second set of values with a complex conjugate version of the second set of values;

- an analyzer being configured for determining the speech fundamental frequency estimate on the basis of the values of the first power density spectrum and the values of the second power density spectrum.

[0016] Analogously, according to said first aspect of the invention a method for estimating a speech fundamental frequency is provided, the method using a first set of values and a second set of values, the first set of values being a received frequency domain representation of a first set of time domain signal values within a first time interval and the second set of values being a received frequency domain representation of a second set of time domain signal values within a second time interval, the second time interval being later than and offset from the first time interval, the method for estimating the speech fundamental frequency comprising the steps of:

- storing a version of the first set of values and providing values of a first power density spectrum by multiplying the stored version of the first set of values with a compley conjugate version of the second set of values;

- providing values of a second power density spectrum by multiplying a version of the second set of values with a complex conjugate version of the second set of values ;

- determining the speech fundamental frequency estimate on the basis of the values of the first power density spectrum and the values of the second power density spectrum.

[0017] This first aspect of the invention is based on the finding that by utilizing the first and second sets of values, which originate from sets of a time domain signal values in the time intervals which are offset from each other, results in a total analyzed signal portion which is a larger than just one single signal portion, for example the first or the second time intervals. Expressed in other words it is now possible to analyze a timely longer signal portion by means of existing (short) time-frequency-transformed signals without the need to provide a new time-frequency-transform just for the estimation of the speech fundamental frequency. However, it is exactly the combination of a given first and second set of values which enables such a timely longer analysis interval, that is the calculation of the first spectrum from the first and second sets of values and the second spectrum from only the second set of values. Thus, the first spectrum represents the spectrum over the longer time interval whereas the second spectrum serves the purpose to determine the characteristics of the second set of values in order to compensate errors in the first spectrum. Therefore it is necessary not only to calculate the first spectrum but also to calculate the second spectrum.

[0018] The approach according to the first aspect of the invention provides the advantage that a signal given in a time-frequency-transformed version (provided for other applications than speech fundamental frequency estimation) can still be used also for speech fundamental frequency estimation (even in the case the time-frequency-transformed version of the signal would normally be not appropriate for providing a precise speech fundamental frequency estimation).

[0019] According to a second aspect of the invention a speech fundamental frequency estimator is provided which is configured for receiving a set of values, the set of values being a frequency domain representation of a set of time domain signal values within a time interval, the speech fundamental frequency estimator comprising:

- a power density spectrum calculator being configured for providing values of a power density spectrum by multiplying a version of the set of values with a complex conjugate version of the set of values, wherein the power density spectrum calculator is configured for determining an estimate of the power density spectrum of background noise and for determining a noise suppression factor on the basis of said power density spectrum of background noise;

- an analyzer being configured for multiplying the power density spectrum with said noise suppression factor and for performing a frequency-time-transform of the multiplied values of the power density spectrum in order to obtain a set of correlation function values, wherein the analyzer is furthermore configured for determining the speech fundamental frequency estimate on the basis of the set of correlation function values.

[0020] Analogously, according to the second aspect of the present invention a method for estimating a speech fundamental frequency is provided, the method being configured for receiving a set of values, the set of values being a frequency domain representation of a set of time domain signal values within a time interval, the method comprising the steps of:

- providing values of a power density spectrum by multiplying a version of the set of values with a complex conjugate version of the set of values;
- determining an estimate of the power density spectrum of background noise and determining a noise suppression factor on the basis of said power density spectrum of background noise and the power density sprectrum of the input signal;
- multiplying the power density spectrum with said noise suppression factor;
- performing a frequency-time-transform of the multiplied values of the power density spectrum in order to obtain a set of correlation function values; and
- determining the speech fundamental frequency estimate on the basis of the set of correlation function values.

[0021] The second aspect of the present invention is based on the finding that a significant improvement in the preciseness of speech fundamental frequency estimation can be realized when background noise is adequately compensated. This is especially the case in a scenario where in speech pauses erroneous detections of speech occur which then falsify the detected result and, in consequence, decrease the reliability of the detected speech fundamental frequency.

[0022] The second aspect of the present invention thus provides the advantage that by simple means, for example a pause detector or just a further analysis of the already existing signal frames a significant improvement in preciseness and reliability of the estimated speech fundamental frequency can be obtained.

[0023] According to a further aspect of the present invention the speech fundamental frequency estimator is characterized in that the first power density spectrum calculator is configured for multiplying versions of the sets of values which represent sets of time domain signal values having overlapping time intervals. This provides the advantage that by multiplying said sets of values, which represent portions of overlapping and therefore consecutive time intervals, a signal in a total time interval can be analyzed, in which a low fundamental frequency can be reliably estimated in given short time signal portions.

[0024] Furthermore, the speech fundamental frequency estimator according to another aspect of the present invention is characterized in that the first power density spectrum calculator is configured for multiplying versions of the sets of values which represent time domain signal values having time intervals overlapping in least 25 percent. This provides the possibility that the speech fundamental frequency estimate can be surely determined as the first and second sets of values belonged to time domain signal values which have a sufficiently overlapping a interval structure. Therefore, due to the sufficient overlap of both time intervals, such an estimation can be considered to be an estimation over the "longer" time interval.

[0025] According to a further aspect of the present invention the speech fundamental frequency estimator is characterized in that the second power density spectrum calculator is configured for providing a conjugate complex version of the second set of values to the first power density spectrum calculator and wherein the first power density spectrum calculator is configured for using the provided conjugate complex version of the second set of values as the version with which the stored version of the first set of values is to be multiplied. This provides the advantage that a complex conjugate version of one of the sets of values has to be calculated only once such that the numerical or computational effort can be reduced.

[0026] In another embodiment of the present invention the speech fundamental frequency estimator is characterized in that the analyzer is configured for performing a first frequency-time-transform of the first power density spectrum in order to obtain a first set of correlation function values and for performing a second frequency-time-transform of the second power density spectrum in order to obtain a second set of correlation function values, wherein the analyzer is furthermore configured for determining a set of normalization values and a set of weighting values from the second power density spectrum and for using the set of normalization values and the set of weighting values in the first and second frequency-time-transform and wherein the analyzer is furthermore configured for determining the speech fundamental frequency estimate on the basis of the first and second sets of correlation function values. This provides the advantage that, on one hand, the short-time envelope can be eliminated and, on the other hand, it is possible to increase the attenuation with rising frequency. Herewith typical characteristics of the speech, especially the speech fundamental frequency structure in the low frequency rage can be adequately be dealt with.

[0027] Also, the speech fundamental frequency estimator according to a further embodiment can be characterized in that the analyzer further comprises a compensator being configured for adaptively compensating the values of the first set of correlation function values by a correction factor being based on a value of the second set of correlation function values and wherein the analyzer is furthermore configured for determining the speech fundamental frequency estimate on the basis of the compensated first set of correlation function values and the second set of correlation function values. Providing such an adaptive compensation control provides the advantage that it is now possible to correct error terms in the cross correlation function as to compensate for example undesired amplitudes which occur at the distinct offsets.

[0028] According to another embodiment the speech fundamental frequency estimator can be characterized in that the compensator is configured for multiplying the second set of correlation function values by a lower bounded quotient between a value of the first set of correlation function values and a value of the second set of correlation function values in order to obtain said compensated first set of correlation function values. Such a configuration of the speech fundamental frequency estimator makes sure that a relation between the cross correlation function and the autocorrelation function does not decrease below a minimal value which, in turn, improves the robustness of speech fundamental frequency estimation.

[0029] Furthermore, it is also possible according to another embodiment of the present invention that the speech fundamental frequency estimator is characterized in that the analyzer is configured for combining the compensated first set of correlation function values and the second set of correlation function values in order to obtain an extended set of correlation function values, wherein the values of the extended set of correlation function values assume corresponding values from the compensated first set of correlation function values, the second set of correlation function values or values between the compensated first set of correlation function values and the second set of correlation function values and wherein the analyzer is furthermore configured for determining the speech fundamental frequency estimate on the basis of said extended set of correlation function values. This provides the advantage that the extended set of correlation function values comprises now information from the first as well as the second set of correlation function values such that an estimation of the speech fundamental frequency can be based on the information comprised in the first and second time interval as well as a correction of possible errors is also possible by the information of the second time interval. Furthermore, it is also possible to perform a weighting of the values of the first set of correlation function values in contrast to the values of the second set of correlation function values in order to take into account the influence of an offset between the first set of correlation function values (respectively the compensated set of correlation function values) and the second set of correlation function values.

[0030] In a further embodiment the speech fundamental frequency estimator is characterized in that the analyzer is configured for determining the speech fundamental frequency estimate by searching the index of a maximum value from the extended set of correlation function values within a predetermined number of indices of the values of the extended set of correlation values, from the first or second set of correlation function values within a predetermined number of indices of values of the first respectively second set of correlation function values or from the compensated first set of correlation function values within the predetermined number of indices of values of the compensated first set of correlation function values and wherein the analyzer is furthermore configured for determining the speech fundamental frequency estimate as the product of a sampling frequency and a reciprocal value of said searched index.

[0031] According to a further embodiment, the speech fundamental frequency is characterized in that the analyzer is furthermore configured for determining a reliability factor for the determined speech fundamental frequency estimate and for blocking an output of the determined speech fundamental frequency estimate in the case the determined reliability factor for the determined speech fundamental frequency estimate is below said predetermined reliability factor. Such a configuration improves the reliability of the estimated speech fundamental frequency.

[0032] Additionally, in a special embodiment the speech fundamental frequency estimator can be characterized in that the analyzer is furthermore configured for determining said reliability factor by dividing the maximum value at said searched index by the first value of the extended set of correlation function values or, respectively the first, the compensated first or second set of correlation function values. This provides the advantage that the reliability factor is only dependent on the scenario in which the speech fundamental frequency estimator is used and not on just a predefined factor which might be too rough in some situations.

[0033] Furthermore, the speech fundamental frequency estimator can be characterized in that the second power density spectrum calculator is configured for determining an estimate of the power density spectrum of background noise and for determining a noise suppression factor on the basis of said power density spectrum of background noise, and wherein the analyzer is configured for multiplying the first and second power density spectrum with said noise suppression factor prior to the frequency-time-transform of the first respectively second power density spectrum. This provides the advantage that an additional improvement can be realized as then erroneous detections in speech pauses can be avoided, which, in turn, improve the reliability of the estimated speech fundamental frequency estimate.

[0034] Especially the speech fundamental frequency estimator can be characterized in that the second power density spectrum calculator is configured for determining the noise suppression factor as the maximum of a predetermined maximum suppression coefficient and a term being dependent on a quotient of the estimate of the power density spectrum of background noise and the second power density spectrum. This makes sure, that a minimum suppression factor is used and thus an effective suppression of background noise is accomplished.

[0035] In a further embodiment of the present invention the speech fundamental frequency estimator can be characterized in that the second power density spectrum calculator is configured for determining the estimate of the power density spectrum of background noise in speech pauses or for determining the estimate of the power density spectrum of background noise from a segment-wise estimation of the minima of the power of a differential signal. This provides an efficient and numerically simple way of determining the estimate of the power density spectrum of background noise.

[0036] In particular, the speech fundamental frequency estimator can be characterized in that the noise suppression factor is defined by

$$V\left(e^{j\Omega_\mu}, n\right) = \max\left\{V_0,\, 1 - \beta\frac{\widehat{S}_{nn}(\Omega_\mu, n)}{\widehat{S}_{yy}(\Omega_\mu, n)}\right\}$$

wherein $\hat{S}_{nn}(\Omega_\mu,n)$ denotes the estimate of the power density spectrum of the background noise, $\hat{S}_{yy}(\Omega_\mu,n)$ denotes the second power density spectrum of the input signal, V_0 denotes a predefined maximum attenuation factor and β denotes a value for overestimating the power density spectrum of the background noise.

[0037] In a further embodiment of the present invention the speech fundamental frequency estimator can be characterized in that the analyzer is furthermore configured for reestimating the speech fundamental frequency estimate in the case the determined speech fundamental frequency estimate is below the predefined frequency value wherein the analyzer is configured for performing the reestimation by searching a further index of a further maximum value of the extended set of correlation function values, the first or second set of correlation function values or the compensated first set of correlation function values within a further number of values of said sets of correlation function values and for outputing a product of a sampling frequency and a reciprocal value of said further index as the determined speech fundamental frequency estimate. This provides a further improvement of the speech fundamental frequency especially in the case when the determined estimate is below said predefined frequency (which means that the estimate may probably not as reliable as actually wanted).

[0038] Especially the speech fundamental frequency estimator can be characterized in that the analyzer is configured for searching said index of said further maximum value using a number of values k of said sets of correlation function values which is defined by

$$\frac{f_s}{f_{\mathrm{p,max}}} \le k < \frac{f_s}{2f_{\mathrm{p}}(n)} + k_0$$

wherein k denotes the number of values of said sets of correlation function values, $f_p(n)$ denotes the previously determined speech fundamental frequency estimate, $f_{p,max}$ denotes a predefined value of a maximal possible speech fundamental frequency, f_s denotes a sampling frequency and k_0 denotes a constant which enables the search of a maximum slightly above $k = \frac{f_s}{2f_p(n)}$. Such a use of the doubled speech fundamental frequency estimate from a previous estimation broadens the region to be searched and thus strengthens the reliability and preciseness of the outputted estimate.

[0039] Also, in another embodiment of the present invention the speech fundamental frequency estimator can be characterized in that the analyzer is configured for outputting said product as the predetermined speech fundamental frequency estimate only in the case the value of the autocorrelation function at the further index is larger than 60 percent of the value of the autocorrelation function at the previously searched maximal index as well as a value of the extended set of correlation function values at said further index is larger than a previously defined amplitude value. This further strengthens the validity of the outputted speech fundamental frequency estimate as before outputting the result two separate conditions have to be fulfilled.

[0040] Additionally the speech fundamental frequency estimator in a further embodiment can be characterized in that the analyzer is configured for modifying a speech fundamental period corresponding to said determined speech fundamental frequency estimate by an interpolation correction term prior of outputting a modified speech fundamental frequency estimate, wherein said interpolation correction term is dependent on values of said first or second set of correlation function values, of said extended set of correlation function values or said compensated first set of correlation function values, respectively. Such an interpolation approach provides the advantage that the error terms resulting from the use of a discrete time-frequency-transform respectively a frequency-time-transform can be reduced by a processing of the signals after the inverse transform has been performed.

[0041] In a further embodiment of the present invention the speech fundamental frequency estimator can be characterized by a frequency domain filtering unit being configured for receiving the frequency domain versions of the first and second set of time domain signal values, for frequency domain filtering said frequency domain versions in order to obtain said first and second sets of values, respectively, and for providing said first and second sets of values to the first and second power density spectrum calculator respectively. Such a pre-processing of the received signals provides the advantage that a pre-processed version of the input signal significantly increases the reliability and preciseness of the estimation in contrast to an embodiment of the invention in which no pre-processing is performed. However the computational or numerical burden for this is relatively low, especially if the filter has a little number of coefficients.

[0042] In a further embodiment of the present invention the speech fundamental frequency estimator can be characterized in that the frequency domain filtering unit is configured for filtering only frequencies below a predefined limiting frequency. This relaxes a computational burden as only the parts of the spectrum are filtered which are of the most importance for a reliable estimation of very low speech fundamental frequencies.

[0043] Furthermore, in another embodiment the speech fundamental frequency estimator can be characterized in that the frequency domain filtering unit is configured for delaying values of said frequency domain versions being above said predefined limiting frequency. This compensates a delay which might be introduced in a signal flow path for filtering signals having a frequency below said limiting frequency.

[0044] The above mentioned aspects and modifications according to the first aspect of the present invention can also be implemented in corresponding methods where the advantages mentioned above come into effect in an analogous manner.

[0045] Furthermore, the invention can also be implemented as a computer program having a program code for performing the inventive method, when the computer program runs on a computer.

[0046] In an embodiment of the present invention focusing the previously mentioned second aspect the speech fundamental frequency estimator can be characterized in that the power density spectrum calculator is configured for determining the noise suppression factor as the maximum of a predetermined maximum suppression coefficient and a term being dependent on a quotient of the estimate of the power density spectrum of background noise and the second power density spectrum. This provides the advantage that an additional improvement can be realized as then erroneous detections in speech pauses can be avoided which, in turn, improves the reliability of the estimated speech fundamental frequency. Also, it can be made sure that the noise suppression factor is always above a predefined value.

[0047] Further, the present invention according to the second aspect may comprise a speech fundamental frequency estimator being characterized in that the power density spectrum calculator is configured for determining the estimate of the power density spectrum of background noise in speech pauses or for determining the estimate of the power density spectrum of background noise from a segment-wise estimation of the minima of the power of a differential signal. This makes sure, that a minimum suppression factor is used and thus an effective suppression of background noise is accomplished.

[0048] Furthermore, the speech fundamental frequency according to a further embodiment may be characterized in that the noise suppression factor is defined by

$$V\left(e^{j\Omega_\mu}, n\right) = \max\left\{V_0,\ 1 - \beta\frac{\hat{S}_{nn}(\Omega_\mu, n)}{\hat{S}_{yy}(\Omega_\mu, n)}\right\}$$

wherein $\hat{S}_{nn}(\Omega_\mu, n)$ denotes the estimate of the power density spectrum of the background noise, $\hat{S}_{yy}(\Omega_\mu, n)$ denotes the second power density spectrum of the input signal, V_0 denotes a predefined maximum attenuation factor and β denotes a value for overestimating the power density spectrum of the background noise.

[0049] The above mentioned aspects and modifications according to the second aspect of the present invention can also be implemented in corresponding methods where the advantages mentioned above come into effect in an analogous manner.

[0050] Furthermore, the invention according to the second aspect can also be implemented as a computer program having a program code for performing the inventive method, when the computer program runs on a computer.

[0051] Additional features and advantages of the present invention will become more readily appreciated from the following detailed description of preferred or advantageous embodiments with reference to the accompanying drawings, in which

Figure 1 shows a time-frequency-analysis of a speech signal;

Figure 2 shows a block diagram of a multi-rate system for speech recognition having a speech fundamental frequency estimation;

Figure 3 shows a block diagram of an analysis system for speech recognition having a speech fundamental frequency estimation;

Figure 4 shows a block diagram of a method and a system for speech fundamental frequency estimation;

Figure 5 shows an autocorrelation- and time-frequency-analysis of sinusoidal signals with varying frequency distances from 300 Hz to 60 Hz;

Figure 6 shows a block diagram of a system respectively method for a speech fundamental frequency estimation with spectral refinement;

Figure 7 shows a block diagram of a system for speech fundamental frequency estimation with spectral refinement in the lower frequency band from 0 Hz to 1000 Hz;

Figure 8 shows in the upper section the analysis of the autocorrelation and in the lower section the time-frequency analysis of sinusoidal signals with varying frequency distances from 300 Hz to 60 Hz. The analyses have been performed with a previous spectral refinement;

Figure 9A shows a block diagram of an embodiment of the inventive speech fundamental frequency estimator;

Figure 9B shows a flow diagram of an embodiment of the inventive method for estimating the speech fundamental frequency estimate;

Figure 10 shows diagrams with results of the speech fundamental frequency estimation with a spectral refinement as a time-frequency-analysis with and without noise reduction;

Figure 11A shows a block diagram of another embodiment of the inventive speech fundamental frequency estimator;

Figure 11B shows a flow diagram of another embodiment of the inventive method for estimating the speech fundamental frequency estimate;

Figure 12 shows a block diagram of a method respectively system for speech fundamental frequency estimation with additional consideration of a passed subband input vector and spectral refinement;

Figure 13 shows in the left section the analysis of the autocorrelation function of a speech fundamental frequency at about 270 Hz. In the right section the analysis of a low speech fundamental frequency of about 60 Hz is shown;

Figure 14 shows in the upper section the analysis of an extended autocorrelation and in the lower section the time-frequency-analysis of several sinusoidal signals with varying frequency distances from 300 Hz to 60 Hz. Additionally a spectral refinement had been performed in a lower frequency range;

Figure 15 shows in the upper section the time-frequency-analysis of speech a signal with additional post-processing and in the lower section the time-frequency-analysis of a speech signal without additional post-processing;

Figure 16 shows in the upper section the time-frequency analysis of several sinusoidal signals of equal amplitude with varying frequency distance (partial section of the signal). In the lower section the time-frequency-analysis of a speech signal (partial section of the signal) is shown.

[0052] Equal or similar elements may have the same reference numbers in the following description of embodiments of the present invention.

Description of preferred embodiments

[0053] The present invention relies mainly on estimation methods based on autocorrelation function which are described herein in advance for a better understanding. However, some aspects of the present invention are also implemented in the conventional autocorrelation methods such that the description in this section is not to be considered as state of the art.

[0054] In the following it is assumed that the speech signal s(n) will be recorded by a microphone. To this signal background noise n(n) is often superimposed. Consequently, the microphone signal y(n) is composed by local speech s(n) and disturbances n(n):

$$y(n) = s(n) + n(n) \qquad (2)$$

[0055] From this signal the short-time autocorrelation function in the time domain can be estimated in a block-based way according to

$$\hat{r}_{yy}(m,n) = \frac{1}{L}\sum_{k=0}^{L-1} y(n-k)y(n-k+m) \qquad (3)$$

[0056] As this short-time autocorrelation function has to be performed for a quite large region of the autocorrelation offset m, the direct estimation requires too much effort for many applications. As in hands-free- and speech recognition systems in multi-rate-structure nevertheless a subband transform (for example by a DFT) is calculated a approach which requires less effort can be used here. The analysis filter bank of a multi-rate system can be described as follows:

- First the input signal y(n) is portioned in windowed, overlapping frame blocks [see also J. Benesty, S. Makino, J. Chen: Speech Enhancement, Springer, Berlin, Deutschland, 2005; E. Hänsler, G. Schmidt: Acoustic Echo and Noise Control - A Practical Approach, John Wiley & Sons, Hoboken, New Jersey, USA, 2004; E. Hänsler, G. Schmidt: Topics in Acoustic Echo and Noise Control - Selected Methods for the Cancellation of Acoustic Echoes, the Reduction of Background Noise, and Speech Processing, Springer, Berlin, Deutschland, 2006 or P. Vary, R. Martin: Digital Speech Transmission, John Wiley & Sons, Chichester, England, 2006). In dependence of a DFT of order N (which is actually the block length of said DFT), one frame block respectively one signal input vector y(n) is composed as follows:

$$\boldsymbol{y}(n) = \left[y(n), y(n-1), \ldots, y(n-N+1)\right]^{\mathrm{T}} \qquad (4)$$

- each signal input vector y(n) is weighted subsequently by a window function

$$\boldsymbol{h} = \left[h_0, h_1, \ldots, h_{N-1}\right]^{\mathrm{T}} \qquad (5)$$

and

- transformed to the frequency domain by a DFT:

$$Y\left(e^{j\Omega_\mu}, n\right) = \sum_{k=0}^{N-1} y(n-k)\, h_k e^{-j\Omega_\mu k} \qquad (6)$$

[0057] The sampling points μ are hereby located equidistantly in the normalized frequency domain:

$$\Omega_\mu = \frac{2\pi}{N}\mu \quad \text{with } \mu \in \{0,\ldots,N-1\} \qquad (7)$$

[0058] From the short-time spectrum $Y(e^{j\Omega_\mu},n)$ the short-time power density spectrum can be estimated by calculating the square of the absolute value according to the following equation:

$$\widehat{S}_{yy}(\Omega_\mu,n) = \left|Y\left(e^{j\Omega_\mu},n\right)\right|^2 = Y\left(e^{j\Omega_\mu},n\right)Y^*\left(e^{j\Omega_\mu},n\right) \qquad (8)$$

[0059] The thus determined power density spectrum $\hat{S}_{yy}(\Omega_\mu,n)$ from equation 8 is then smoothed in frequency direction and divided by the thus obtained envelope $\overline{S}_{yy}(\Omega_\mu,n)$. Hereby the short-time envelope is removed. The smoothing in frequency direction can be described by

$$\widetilde{S}_{yy}(\Omega_\mu,n) = \begin{cases} \widehat{S}_{yy}(\Omega_\mu,n), & \text{for } \mu = 0, \\ \lambda\,\widetilde{S}_{yy}(\Omega_{\mu-1},n) + (1-\lambda)\,\widehat{S}_{yy}(\Omega_\mu,n), & \text{for } \mu \in \{1,\ldots,N-1\} \end{cases} \qquad (9)$$

respectively

$$\overline{S}_{yy}(\Omega_\mu,n) = \begin{cases} \widetilde{S}_{yy}(\Omega_\mu,n), & \text{for } \mu = N-1, \\ \lambda\,\overline{S}_{yy}(\Omega_{\mu+1},n) + (1-\lambda)\,\widetilde{S}_{yy}(\Omega_\mu,n), & \text{for } \mu \in \{0,\ldots,N-2\} \end{cases} \qquad (10)$$

[0060] Values for the smoothing constant λ are chosen from the range

$$0.3 < \lambda < 0.7 \qquad (11).$$

[0061] Following, a linear weighting of the estimated and normalized power density spectrum is performed:

$$\widehat{S}_{yy,\mathrm{norm}}(\Omega_\mu,n) = \frac{\widehat{S}_{yy}(\Omega_\mu,n)}{\overline{S}_{yy}(\Omega_\mu,n)}\,W\left(e^{j\Omega_\mu}\right) \qquad (12).$$

[0062] The weighting function $W(e^{j\Omega_\mu}, n)$ has been chosen such that the attenuation rises with rising frequency. This choice results from the fact that speech mainly at low frequencies has a speech fundamental frequency structure - which in turn results in an improved estimation of the speech fundamental frequency. In Fig. 4 the functional principle of a method for speech fundamental frequency estimation is shown.

[0063] The autocorrelation function

$$\hat{r}_{yy}(m,n) = \frac{1}{N}\sum_{\mu=0}^{N-1} \widehat{S}_{yy,\mathrm{norm}}(\Omega_\mu, n)\, e^{j\frac{2\pi}{N}\mu m} \qquad (13)$$

is determined by an inverse transform of the normalized and weighted power density spectrum from equation 12. The autocorrelation function $\hat{r}_{yy}(m,n)$ is used in order to estimate the speech fundamental frequency f_p (n). The index m describes herein the autocorrelation offset and the index n describes the present frame (under analysis). For each a single frame the preliminary speech fundamental frequency $f'_p(n)$ can be determined by a search of the maximum in a selected range of indices, for example 30 ≤ m ≤ 100. The speech fundamental frequency is then determined as the reciprocal of value of the index at which the maximum of the autocorrelation has occurred (in view to the sampling frequency f_s):

$$f'_\mathrm{p}(n) = \frac{f_\mathrm{s}}{\tau_\mathrm{p}(n)} \qquad (14)$$

with

$$\tau_\mathrm{p}(n) = \underset{30 \le m \le 100}{\operatorname{argmax}} \left\{ \hat{r}_{yy}(m,n) \right\} \qquad (15)$$

[0064] Furthermore a reliability $p_{f_p}(n)$ of the estimated speech fundamental frequency is determined. Therefore the value of the normalized autocorrelation at the maximum point, (i.e. the index where the autocorrelation function becomes maximal) is used:

$$p_{f_\mathrm{p}}(n) = \frac{\hat{r}_{yy}(\tau_\mathrm{p}(n), n)}{\hat{r}_{yy}(0,n)} \qquad (16)$$

[0065] Large values, that are values in the proximity to one, indicate a very sure detection - small values indicate a doubtful detection. For this reason a detection only takes place for values of the normalized autocorrelation function which are larger than p_0 (which is taken as a predefined threshold value):

$$f_p(n) = \begin{cases} f'_p(n) & \mathit{for} \quad p_{f_p}(n) > p_0 \\ \mathit{not\ detectabale,} & \mathit{else} \end{cases} \qquad (17)$$

[0066] A threshold value of $p_0 \in [0.2, 0.3]$ has turned out to be favourable. The value of the normalized autocorrelation at the location $\tau_p(n)$ can be of large significance as reliability information, for example for a speech signal reconstruction. Hereby the desired value of the speech fundamental frequency can be either slowly or quickly traced, dependent on how sure a speech fundamental frequency can be estimated.

[0067] Finally the inventive method proposed here is further presented in more detail by an example. Therefore 10 sinusoidal signals of equal amplitude are summed up. The frequencies of the sinusoidal signals have been chosen equidistantly. At the beginning of the signal a fundamental frequency of 300 Hz has been chosen, subsequently this frequency has been decreased linearly over the time to an end value of 60 Hz. In the upper diagram of Fig. 5 the development of the normalized autocorrelation vectors is shown and in the lower diagram of Fig. 5 a time-frequency-analysis of the corresponding input signals y(n) is shown.

[0068] For the analysis a DFT of order N = 256 (= DFT block length), a sampling frequency of f_s = 11025 Hz and the frame offset of r = 64 is used. The analysis of the autocorrelation $\hat{r}_{yy}(m,n)$ has been performed in the range between m=40 to m=128. Detection results have been considered to be well it if the reliability information $p_{f_p}(n)$ is larger than p_0 = 0.2. Finally the time-frequency analysis was considered only in the interesting frequency range up to 1000 Hz.

[0069] In the analysis of the autocorrelation it can be recognized that the speech fundamiental frequency up to an offset of about m=95 can be estimated surely - this corresponds to a speech fundamental frequency of about $f_p(n)$ = 120 Hz (at a sampling frequency of f_s=11025 Hz). The graph of this detection with decreasing frequency can also be seen in the time-frequency-analysis up to about t=3.8 s. However, if speech fundamental frequency is below $f_p(n)$ = 120 Hz (which is often the case with men having a low speech fundamental frequency) these speech fundamental frequency can not be determined in a reliable way.

[0070] Contrary to the approaches mentioned in the previous description of the invention the approach disclosed subsequently has the following further advantages:

- a sure and reliable estimation can also be performed for a very low voices;
- a better robustness in environments with background noise can be reached; and
- the speech fundamental frequency can be determined with a significantly higher degree of precision.

[0071] Firstly, a method for estimating the speech fundamental frequency having an additional spectral refinement is described in more detail and it is shown how the detection robustness can be increased by a noise reduction which is integrated in the estimation method (no pre-processing). Following an additional part of the method is presented which enables to also detect a very low speech fundamental frequencies by an additional delay correction structure. Finally approaches for adaptively post-processing and interpolating are disclosed which enable an error correction respectively an improvement of the preciseness of the speech fundamental frequency. However it has to be mentioned here that all the disclosed aspects can also be used independently such that the present invention does not only work if all the aspects mentioned above are implemented. For example the spectral refinement can be used without using the post-processing or the interpolation or the approach having the additional delay correction structure can be used without using the spectral refinement approach. However all the individual aspects commonly contribute to a much improved estimation of the speech fundamental frequency and shall be described herein as an embodiment.

Speech fundamental frequency estimation with spectral refinement

[0072] In the preceding section it has been shown that a speech fundamental frequency which is below 120 Hz can not be estimated. In the following an approach is presented which solves the described problem.

[0073] Additionally to the already mentioned method according to the state of the art the newly proposed method uses an additional spectral refinement of the input spectrum $Y(e^{j\Omega_\mu}, n)$. The functional principle of this approach is disclosed in Fig. 6. The short-time spectrum $Y(e^{j\Omega_\mu},n)$ is firstly filtered subband-wise by an FIR-filter (FIR = finite impulse response). Such a filtering serves the purpose to perform a more precise spectral resolution of the input spectrum $Y(e^{j\Omega_\mu}, n)$.

[0074] It was shown in Patent Application No. EP 06024940.6 that a spectral refinement within one subband can be reached by a short FIR-filter, respectively, how the individual filter coefficients have to be determined. The disclosure of Patent Application No. EP 06024940.6 is incorporated herein in by reference its entirety. The FIR-filter used for the μ-th subband can be described as follows:

$$\boldsymbol{g}_\mu = \left[g_{\mu,0},\, g_{\mu,1},\, \ldots,\, g_{\mu,M-1}\right]^{\mathrm{T}} \qquad (18)$$

[0075] The parameter μ denotes herein the μ-th frequency sampling point of a short-time spectrum $\tilde{Y}(e^{j\Omega_\mu},n)$ having a higher resolution and the parameter M denotes the order of the used FIR-filters. A memory length M of the short FIR-filter is chosen between 3 and 5. For the frequency subbands of interest the spectral refinement finally can be determined as follows:

$$\tilde{Y}(e^{j\Omega_\mu},n) \;=\; g_{\mu,0}\,Y(e^{j\Omega_\mu},n)+\cdots+g_{\mu,M-1}\,Y(e^{j\Omega_\mu},n-(M-1)\cdot r) \qquad (19)$$

[0076] A spectral refinement in the whole frequency range is not necessary for speech signals. Usually the speech fundamental frequency structure is only present in the lower frequency range that means it is sufficient to perform the refinement up to, for example, 1000 Hz. Above this threshold it is possible to only introduce a delay of (M-1)/2 samples (down-sampled). The numerical effort necessary for such a refinement can thus be kept low. In Fig. 7 the analysis-synthesis-system with additional calculation of the spectral refinement in a low frequency range is shown.

[0077] However, it has to be mentioned that by the calculation of a spectral refinement a low delay is introduced into the signal path. A detailed derivation of this part of the new approach is explained in more detail in Patent Application No. EP 06024940.6 which is incorporated herein in its entiery.

[0078] Subsequently the determination of the speech fundamental frequency can be performed analogously to the way as already disclosed in the previously mentioned description. However, the refined short-time spectrum $\tilde{Y}(e^{j\Omega_\mu},n)$ is now used in order to calculate the estimated and refined power density spectrum $\hat{S}_{\tilde{y}\tilde{y}}(\Omega_\mu,n)$ according to the following equation:

$$\hat{S}_{\tilde{y}\tilde{y}}(\Omega_\mu,n) = \tilde{Y}(e^{j\Omega_\mu},n)\tilde{Y}^*(e^{j\Omega_\mu},n) = \left|\tilde{Y}(e^{j\Omega_\mu},n)\right|^2 \qquad (20)$$

[0079] Following the power density spectrum $\hat{S}_{\tilde{y}\tilde{y}}(\Omega_\mu,n)$ is also smoothed, weighted and the autocorrelation function $\hat{r}_{\tilde{y}\tilde{y}}(m,n)$ for the estimation of the speech fundamental frequency is determined. In order to calculate said power density spectrum an approach corresponding to equations 9 to 17 can be used.

[0080] In Fig. 8 the analysis of autocorrelation as well as the time-frequency-analysis with spectral refinement is shown. For the analyses the same parameters as previously mentioned have been used -namely a DFT of order N=256, a sampling of frequency f_s= 11025 Hz, a frame offset of r=64 and a detection of threshold p_0 = 0.2. Furthermore as test signal the same combination from sinusoidal signals have been used which have a varying frequency distance of 300 Hz to 60 Hz. The black graph in the upper diagram of Fig. 8 as well as the white graph in the lower diagram of Fig. 8 show the estimated pitch period duration, respectively; the estimate of speech fundamental frequency when using the spectral refinement approach.

[0081] A comparison of Fig. 5 and 8 shows very clearly that the spectral refinement provides the possibility of a far better detection of the speech fundamental frequency. Very desirable is the fact that the sure and reliable detection rises up to an offset of m=N/2 =128 - this corresponds to a speech fundamental frequency of about 90 Hz. At lower frequencies f_p < 90 Hz several detection errors occur. Finally it has to be mentioned that in many applications it is only of interest whether a speech fundamental frequency is present or not - an exact speech fundamental frequency would be of minor importance. Just in these application scenarios the previously presented approach would provide significant advantages.

[0082] In the following it will be the aim to present a new approach which works robustly in terms of erroneous estimations at very low speech fundamental frequencies. Additionally it is shown in the following section how noise reduction can be advantageously incorporated into the presently known method.

Speech fundamental frequency estimation with noise suppression

[0083] Fig. 9A shows a block diagram of an embodiment of a speech fundamental frequency estimator 900. The speech fundamental frequency estimator 900 comprises a power density spectrum calculator 902 and an analyzer 904. The power density spectrum calculator 902 has 2 inputs, one for receiving a set of values and one for receiving background noise information. The set of values $\tilde{Y}_1$, is a frequency-domain representation of a set of a time domain signal values y_1 in a time interval t_1. The background noise information can for example be determined in speech pauses in which only a noise signal and no speech signal is provided to the power density spectrum calculator 902. The power density spectrum calculator 902 has 2 outputs, one for outputting a noise suppression factor V($e^{j\Omega_\mu}$,n) and one for outputting values of a power density spectrum. The analyzer 904 has 2 inputs for receiving both of the outputs of the power density spectrum calculator 902. The analyzer 904 has a furthermore one output for outputting the determined speech fundamental frequency f_p(n).

[0084] The function of the speech fundamental frequency estimator 900 shall be described in more detail with reference to Fig. 9B. In Fig. 9B a flow diagram of a method for estimating the speech fundamental frequency is disclosed. The method 940 comprises a first step 950 in which a power density spectrum is provided by multiplying a version of the set of values $\tilde{Y}_2$ with a complex conjugate version of the second set of values. In parallel (or in series) in a second step 952 an estimate of a power density spectrum of background noise is determined. In this step 952 of determining the estimate of a power density spectrum of background noise the background noise information is used which may originate for example from a speech pause detector or other means which provide only information about the background noise in the absence of speech. In a third step 954 a noise suppression factor is determined which is explained in more detail below. In a fourth step 956 a multiplication of the power density spectrum with the noise suppression factor V($e^{j\Omega_\mu}$,n) is performed before in a fifth step 958 a frequency-time-transform is accomplished. Subsequently in a sixth step 960 speech fundamental frequency is determined from the frequency-time-transformed signal resulting in step 958.

[0085] Such an approach provides the advantage that by considering background noise information the detection preciseness as well as in detection robustness can be improved as for example in speech pauses when only background noise occurs no speech fundamental frequency shall be estimated. Thus, the reliability of an estimated speech fundamental frequency can be significantly improved. This results from the fact that the erroneous detections of speech fundamental frequencies in speech pauses can be avoided. Furthermore the multiplication of the noise suppression factor with the power density spectrum prior to the frequency-time-transform provides the advantage that such a multiplication in the frequency domain requires very little computational and numerical effort in contrast to a similar combination in time domain. Furthermore it is also possible to additionally considered other calculations or normalizations of the noise-compensated signal prior to said frequency-time-transform.

[0086] To be more precise, methods for the noise reduction are mostly based on modified Wiener-filters which frequency response in the respective frequency intervals is determined by

$$V\left(e^{j\Omega_\mu},n\right) \;=\; \max\left\{V_0,\,1-\beta\frac{\widehat{S}_{nn}(\Omega_\mu,n)}{\widehat{S}_{yy}(\Omega_\mu,n)}\right\} \tag{21}$$

(see also S. F. Boll: Suppression of Acoustic Noise in Speech Using Spectral Subtraction, IEEE Trans. Acoust. Speech Signal Process., Vol. 27, Nr. 2, Seiten 113-120, 1979; E. Hänsler: Statistische Signale - Grundlagen und Anwendungen, Springer, Berlin, Deutschland, 2001 or T. Haulick, K. Linhard: Noise Subtraction with Parametric Recursive Gain Curves, Proceed. of the European Conf. on Speech Communications and Technology, Vol. 6, pages 2611-2614, 1999). The value $\hat{S}nn(\Omega_\mu,n)$ denotes an estimation of the auto power density spectrum of a disturbance (background noise), V_0 describes a maximal attenuation and the parameter β is used for overestimating the power density spectrum of the disturbance. Because of the fact that the disturbance can be considered to be non-stationary a short-time estimation value has to be used for this disturbance value. However, signal and disturbance are available only as a sum in the microphone signal y(n). The estimation of the power density spectrum of the background noise can be obtained in two different ways, firstly the power of the microphone signal can be estimated in speech pauses - which requires a speech pause detector - or, secondly, that an estimated value for the power of the disturbance can be determined from the segment-wise estimated minima of the power of the difference signal. As the noise estimation is not the main focus in this patent application other details shall not be explained here; however reference is made to P. Vary, R. Martin: Digital Speech Transmission, John Wiley & Sons, Chichester, England, 2006 which disclosure is incorporated herein in its entirety by reference.

[0087] Normally, noise reductions are used as a pre-processing stage for a speech fundamental frequency estimation that is instead of the input subband signals $Y(e^{j\Omega_\mu},n)$ the noise reduced signals $Y(e^{j\Omega_\mu}, n)\cdot V(e^{j\Omega_\mu}, n)$ are processed. The present approach follows a similar way that means that firstly a noise-reduced power density spectrum (see equation 12) respectively after a subsequent spectral refinement is determined according to the following equation:

$$\hat{S}_{\tilde{y}\tilde{y},\mathrm{norm,g}}(\Omega_\mu, n) = \frac{\hat{S}_{\tilde{y}\tilde{y}}(\Omega_\mu, n)}{\overline{S}_{\tilde{y}\tilde{y}}(\Omega_\mu, n)} W\left(e^{j\Omega_\mu}\right) \cdot V\left(e^{j\Omega_\mu}, n\right) \tag{22}$$

[0088] For detection the inverse transform is then calculated as follows:

$$\hat{r}_{yy,\mathrm{g}}(m,n) = \frac{1}{N} \sum_{\mu=0}^{N-1} \hat{S}_{\tilde{y}\tilde{y},\mathrm{norm,g}}(\Omega_\mu, n)\, e^{j\frac{2\pi}{N}\mu m} \tag{23}$$

[0089] As standardization factor the value $\hat{r}_{yy}(0, n)$ from the equation 16 is again used which is the standardization value of the autocorrelation including noise. This results in the following modified detection

$$p_{f_\mathrm{p}}(n) = \frac{\hat{r}_{yy,\mathrm{g}}(\tau_\mathrm{p}(n), n)}{\hat{r}_{yy}(0,n)} \tag{24}$$

[0090] As a result a more robust detection in speech pauses is obtained. In order to more clearly show this effect Fig. 10 shows results of the speech fundamental frequency estimation with spectral refinement in terms of time-frequency-analysis with and without noise reduction. All parameters of the methods have been identical to the previously described parameters. As can be seen very clearly erroneous detections (denoted by black ellipses in the upper diagram of Fig. 10) can be suppressed in the case when the above-mentioned active noise reduction is used. In speech activity passages nearly nothing changes.

Speech fundamental frequency estimation on the basis of a plurality of subband vectors

[0091] In this section a further part of the approach for the inventive speech fundamental frequency estimation is described.

[0092] Fig. 11A shows a block diagram of an embodiment of the inventive speech fundamental frequency estimator 1100. The speech fundamental frequency estimator 1100 comprises a first power density spectrum calculator 1102, a second power density spectrum calculator 1104 and an analyzer 1106. The first power density spectrum calculator 1102 and second power density spectrum calculator 1104 are both fed by a common input of width N, on which subsequently a first set of values $\tilde{Y}_1$ and a second set of values $\tilde{Y}_2$ is provided. Herein, the first set of values $\tilde{Y}_1$ is a frequency domain representation of a first set of time domain signal values y_1 within a first time interval t_1. The second set of values $\tilde{Y}_2$ is a frequency domain representation of a second set of time domain signal values y_2 within a second time interval t_2. In the embodiment as shown in Fig. 11A the first and second time intervals overlap. The first power density spectrum calculator 1102 is configured for storing a version of the first set of values and for providing values of a first power density spectrum $\hat{S}_{\tilde{y}\tilde{y}_d}(\Omega_\mu, n)$ by multiplying the stored version of the first set of values $\tilde{Y}_1$ with a complex conjugate version of the second set of values $\tilde{Y}_2$. The second power density spectrum calculator 1104 is configured for providing values of a second power density spectrum $\hat{S}_{\tilde{y}\tilde{y}}(\Omega_\mu, n)$ by multiplying a version of the second set of values with a complex conjugate version of the second set of values. The analyzer 1106 is configured for receiving the first and second power density spectrums of the first respectively second power density spectrum calculator 1102, 1104 and for determining the speech fundamental frequency estimate $f_p(n)$ on the basis of the values of the first power density spectrum $\hat{S}_{\tilde{y}\tilde{y}_d}(\Omega_\mu, n)$ and the values of the second power density spectrum $\hat{S}_{\tilde{y}\tilde{y}}(\Omega_\mu, n)$.

[0093] Fig. 11B shows the functionality of the speech fundamental frequency estimator as shown in Fig. 11A in more detail. To be more precise, Fig. 11B discloses a method 1140 for estimating the speech fundamental frequency $f_p(n)$. Firstly, first and second sets of values $\tilde{Y}_1$ and $\tilde{Y}_2$ are provided, each of which have the number of N individual values (that is a width of N). In a first step 1150 a version of the first set of values $\tilde{Y}_1$ is stored. In a second step 1152 the stored version of the first set of values $\tilde{Y}_1$ it is multiplied with a version of the second set of values $\tilde{Y}_2$ which are directly fed to the multiplication step without a storing step. The result from the multiplication step 1152 is said first power density spectrum $\hat{S}_{\tilde{y}\tilde{y}_d}(\Omega_\mu, n)$. Parallel to the step of multiplying 1152 a further step of multiplying 1154 is performed in which a versions of the second set of values $\tilde{Y}_2$ are multiplied with each other, which results in the second power density spectrum. In a final step 1156 the speech fundamental frequency estimate $f_p(n)$ is determined.

[0094] The inventive approach as shown in Fig. 11A and 11B has the advantage that it is now possible to estimate lower speech fundamental frequencies as would be possible according to the state of the art. This is mainly due to the fact that (conventional existing) short frequency domain values can be used for a precise speech fundamental frequency estimation as the multiplication in step 1152 with a stored respectively delayed version of a previous set of frequency domain values results in a kind of elongated analysis time interval for estimating the low speech fundamental frequency. However, it is also possible to correct possible errors which might result from the time offset of the first and second time intervals because for the determination of the speech fundamental frequency estimate also the second power density spectrum is used which is based on a multiplication of versions of the second set of values. Therefore the first power density spectrum can be compared with the information resulting from the second power density spectrum such that a kind of normalization can be performed or a detection of possible errors in the first power density spectrum can be recognized and corrected.

[0095] To be more specific, in the previous description it has been shown that a speech fundamental frequency below $f_p(n)<120$ Hz can not be detected correctly anymore. Therefore, in the first approach a subsequent spectral refinement has been applied. However, this spectral refinement provided the possibility for an improvement of the speech fundamental frequency estimation only to about $f_p(n)=90$ Hz. The reason for this threshold can be seen in the fact that in the used DFT of order N a maximal autocorrelation offset of $m = N/2 + 1$ for the analysis of this speech fundamental frequency is possible - this corresponds to a maximally low speech fundamental frequency detection of about 90 Hz. It has been assumed that the used power density spectra, respectively the autocorrelation functions are only real (and not complex) and are furthermore also symmetrically.

[0096] A further inventive idea it can be seen in the fact that not only the present signal frame y(n) is used for the estimation of the speech fundamental frequency but also a signal frame y(n-d) which is a signal frame delayed by d clock cycles. For example the speech fundamental frequency estimation can be significantly improved by utilizing of the present signal frame and a signal of frame delayed by one frame cycle, d = r, with an overlap of 75% - this corresponds to a frame offset of r = 64 and a signal block length of N = 256.

[0097] In Fig. 12 the functional principle of the method for estimating the speech fundamental frequency is shown. Additionally to the already described method the inventive approach uses a cross correlation with the delayed input frame. Firstly it can be seen from Fig. 12 that in addition to the estimated auto power density spectrum $\hat{S}_{\tilde{y}\tilde{y}}(\Omega_\mu, n)$ in the lower path of Fig. 12 also a variant of the cross power density spectrum

$$\hat{S}_{\tilde{y}\tilde{y}_\mathrm{d}}(\Omega_\mu, n) = \tilde{Y}\left(e^{j\Omega_\mu}, n\right)\tilde{Y}^*\left(e^{j\Omega_\mu}, n-d\right) \quad (25)$$

is determined too. For the determination of the cross power density spectrum $\hat{S}_{\tilde{y}\tilde{y}_d}(\Omega_\mu, n)$ the present short-time spectrum $\tilde{Y}(e^{j\Omega_\eta}, n)$ and the delayed short-time spectrum $\tilde{Y}^*(e^{j\Omega_\eta}, n\text{-}d)$ is used. In following only the short-time spectrum delayed by one frame clock, that is d = r, is dealt with further, however also other delays can be used here.

[0098] The thus determined cross power density spectrum is divided by the smoothed auto power density spectrum $\overline{S}_{\tilde{y}\tilde{y}_d}(\Omega_\mu,n)$ and is multiplied with a weighting function as shown below:

$$\widetilde{S}_{\tilde{y}\tilde{y}_\mathrm{d}}(\Omega_\mu,n)=\frac{\widehat{S}_{\tilde{y}\tilde{y}_\mathrm{d}}(\Omega_\mu,n)}{\overline{S}_{\tilde{y}\tilde{y}}(\Omega_\mu,n)}W\left(e^{j\Omega_\mu}\right) \qquad (26)$$

[0099] After a subsequent noise reduction and an inverse transform according to equation 23 the cross-correlation function $\hat{r}_{\tilde{y}\tilde{y},g}(m,n)$ is determined according to equation 13. In the following, the aim will be to determine an extended autocorrelation function $\hat{r}_{\tilde{y}\tilde{y},erw}(k,n)$ of order N/2 + r from the autocorrelation function $\hat{r}_{\tilde{y}\tilde{y}_d,g}(m,n)$ and the cross-correlation function $\hat{r}_{\tilde{y}\tilde{y}_d,g}(m,n)$, each of which having the order N/2. The index k of the term $\hat{r}_{\tilde{y}\tilde{y},erw}(k,n)$ describes herein the offset of the autocorrelation, wherein the following equation is valid:

$$k\in\left\{0,\ldots,\frac{N}{2}+r-1\right\} \qquad (27)$$

[0100] By using an adaptive compensation control it can be tried to correct the error terms of the cross-correlation function $\hat{r}_{\tilde{y}\tilde{y}_d,g}(m,n)$. For this purpose a correction value Δ(m, n) is determined for each time frame in order to compensate, for example the undesired amplitudes which occur at an offset of m=r=64, or respectively, to correct the remaining amplitude values in order to perform a later combination with the autocorrelation function $\hat{r}_{\tilde{y}\tilde{y},g}(m,n)$:

$$\hat{r}_{\tilde{y}\tilde{y}_\mathrm{d},\mathrm{g},\mathrm{mod}}(m,n)=\hat{r}_{\tilde{y}\tilde{y}_\mathrm{d},\mathrm{g}}(m,n)-\Delta(m,n)=\hat{r}_{\tilde{y}\tilde{y}_\mathrm{d},\mathrm{g}}(m,n)-c(n)\,\hat{r}_{\tilde{y}\tilde{y},\mathrm{g}}(m-r,n) \qquad (28)$$

[0101] The adaptive constant c(n) is derived from a relation of the cross correlation function $\hat{r}_{\tilde{y}\tilde{y}_d,g}(m,n)$ at the location m= r and the autocorrelation function $\hat{r}_{\tilde{y}\tilde{y},g}(m,n)$ at the location m=0. In order to perform a robust speech fundamental frequency estimation the relation should not be below a minimum value c_0. Therefore the adaptive parameter c(n) is determined as follows:

$$c(n)=\max\left\{\frac{\hat{r}_{\tilde{y}\tilde{y}_\mathrm{d},\mathrm{g}}(r,n)}{\hat{r}_{\tilde{y}\tilde{y},\mathrm{g}}(0,n)},\ c_0\right\} \qquad (29)$$

[0102] Tests have shown that good results can be obtained in the case the constant c_0 is set to a value of c_0=0.4.

[0103] Following the auto and cross-correlation coefficients of $\hat{r}_{\tilde{y}\tilde{y},g}(m,n)$ and $\hat{r}_{\tilde{y}\tilde{y}_d,g,\mathrm{mod}}(m,n)$ are weighted by a weighting function and are combined as follows:

$$\hat{r}_{\tilde{y}\tilde{y},\mathrm{erw}}(k,n) = \begin{cases} \hat{r}_{\tilde{y}\tilde{y},\mathrm{g}}(k,n), & \text{for } 0 \le k < \frac{N}{2} - r, \\ a(k-r)\hat{r}_{\tilde{y}\tilde{y},\mathrm{g}}(k,n) + \Big(1 - a(k-r)\Big)\hat{r}_{\tilde{y}\tilde{y}_\mathrm{d},\mathrm{g},\mathrm{mod}}(k-r,n), & \text{for } \frac{N}{2} - r \le k < \frac{N}{2}, \\ \hat{r}_{\tilde{y}\tilde{y}_\mathrm{d},\mathrm{g},\mathrm{mod}}(k-r,n), & \text{for } \frac{N}{2} \le k < \frac{N}{2} + r. \end{cases} \tag{30}$$

[0104] Herein the linear function a(m) was chosen such that with an increasing offset m the weight of the coefficients reduces. The thus obtained extended autocorrelation function $\hat{r}_{yy;erw}(k,n)$ is finally used for the estimation of the speech fundamental frequency. In comparison to the methods mentioned before the speech fundamental frequency is determined by a search of the maximum for each single frame in an elongated area - for example in the range $30 \le k \le 180$.

[0105] In order to clarify the functioning of the described method in Fig. 13 two examples for the analysis of the speech fundamental frequency are shown. For this purpose the left section of Fig. 13 discloses the analysis of the speech fundamental frequency at about 270 Hz whereas in the right section of Fig. 13 the analysis of a speech fundamental frequency at about 60 Hz is shown.

[0106] In the first aspect the correlation of the present signal frame with itself (left) and with a proceeding signal frame (right) are shown each, the left and also the right section of Fig. 13. The grey graph denotes in each section the cross correlation function $\hat{r}_{\tilde{y}\tilde{y}_d,g}(m,n)$ before the adaptive compensation control and the dark grey graph denotes the cross correlation function $\hat{r}_{\tilde{y}\tilde{y}_d,g,\mathrm{mod}}(m,n)$ after the adaptive compensation control. It can be well identified that significant error terms - especially the error terms at the location k = r - are corrected by the adaptive compensation control.

[0107] The lower graph in each of both sections of Fig. 13 shows the extended autocorrelation function $\hat{r}_{\tilde{y}\tilde{y},erw}(k,n)$ across an elongated autocorrelation offset which is generated by the composition of both correlation functions $\hat{r}_{\tilde{y}\tilde{y},g}(m,n)$ and $\hat{r}_{\tilde{y}\tilde{y},g,\mathrm{mod}}(m,n)$ respectively by the usage of the equation 30. At a high speech fundamental frequency the corresponding speech fundamental period can be determined and detected quite well using the autocorrelation function $\hat{r}_{\tilde{y}\tilde{y},g}(m,n)$ (left section of Fig. 13). In contrast, with the used low speech fundamental frequency of about 60 Hz the corresponding speech fundamental period can not be determined any longer by the standard autocorrelation $\hat{r}_{\tilde{y}\tilde{y},g}(m,n)$. The right section of Fig. 13 shows in the lower part that by a combination of the correlation of the signal frame with itself and the correlation with a proceeding signal frame the speech fundamental period can still be determined and detected.

[0108] In Fig. 14 the analysis of the extended autocorrelation function $\hat{r}_{\tilde{y}\tilde{y},erw}(k,n)$ is shown when a previous spectral refinement in the low frequent region as well as a time-frequency-analysis of the input signal is used. A comparison with the analyses from the Fig. 5 and 14 indicates that by using the previously described approach significant improvements can be achieved. Through this approach a existing speech fundamental period up to an offset of about k= 125 can still be detected. Moreover no erroneous detections with low speech fundamental frequencies occur. Thus, a sure and reliable estimation can be performed by the described approach down to a speech fundamental frequency of about $f_p(n)$= 60 Hz.

Adaptive post-processing

[0109] At several locations erroneous estimations of the speech fundamental frequency $f_p(n)$ still occur. For these values a half, respectively a third, of the speech fundamental frequency are often estimated. A subsequent post-processing is then preferably used to correct the occurring erroneous detections.

[0110] After estimation of the speech fundamental frequency $f_p(n)$ a test can be made whether this estimate is below a threshold f_k. The post-processing only shall be performed in the case the following condition

$$f_p(n) < f_k \qquad (31)$$

is fulfilled. Values between f_k = 140 Hz and f_k = 160 Hz have been recognized to be suitable in practice. Subsequently a normalized speech fundamental period is estimated by performing a search for the index of the maximum of the autocorrelation function

$$\tilde{\tau}_{\mathrm{p}}(n) = \underset{k}{\operatorname{argmax}}\left\{\hat{r}_{\tilde{y}\tilde{y},\mathrm{erw}}(k,n)\right\} \qquad (32)$$

in a selected range

$$\frac{f_s}{f_{\mathrm{p,max}}} \le k < \frac{f_s}{2f_{\mathrm{p}}(n)} + k_0 \qquad (33)$$

[0111] For the determination of this area the previously determined speech fundamental frequency $f_p(n)$ is firstly doubled. The parameter $f_{p,max}$ in equation 33 is herein a predefined value of a maximal possible speech fundamental frequency. Finally the value k_0 is a constant which makes sure that also a search for a maximum which is slightly above $k = \frac{f_s}{2f_p(n)}$ is allowed.

[0112] In the case the newly determined maximum is higher than 60 percent of the previously determined maximum, that is

$$\tilde{\tau}_{\mathrm{p}}(n) \quad > \quad 0.6\,\tau_{\mathrm{p}}(n) \qquad (34)$$

and in the case also the amplitude of this newly determined maximum is above a predetermined amplitude value

$$\hat{r}_{\tilde{y}\tilde{y},\mathrm{erw}}\left(\tilde{\tau}_{\mathrm{p}}(n), n\right) \quad > \quad \tilde{p}_0 \qquad (35)$$

a correction of the previously estimated speech fundamental frequency is performed according to

$$f_{\mathrm{p}}(n) \;=\; \frac{f_{\mathrm{s}}}{\tilde{\tau}_{\mathrm{p}}(n)}. \qquad (36)$$

[0113] In order to clarify the improvements which result from such a post-processing, Fig. 15 shows a time-frequency-analysis of an input signal, respectively, the detection results of the speech fundamental frequency estimation. In the upper part of Fig. 15 the post-processing was deactivated and at two locations (at 0.7 and at 0.75 seconds) erroneous detections (bisections of frequency) can be observed. Such erroneous detections can be corrected by the post-processing which can be concluded from the lower part of Fig. 15.

Interpolation

[0114] In the application of the approach described up to now it could be observed that only an inaccurate speech fundamental frequency is estimated. In the estimation results stairs-like graphs of the estimated speech fundamental frequency have been generated as can be seen in Fig. 14 for example. Up to now it was only possible to determine the quantized speech fundamental frequency estimate, that means when the exact speech fundamental period is in between two autocorrelation offsets k of the autocorrelation function $\hat{r}_{\tilde{y}\tilde{y},erw}(k,n)$ then a rounding to the nearest autocorrelation offset k is performed in order to determine the estimated speech fundamental period $\tau_p(n)$, respectively $\tilde{\tau}_p(n)$. Therefore quantization errors occur.

[0115] In numerous applications, as for example for a speech signal construction, an exact speech fundamental frequency estimation is of significant importance. One possible approach to solve the described problem is to perform an interpolation of the estimated speech fundamental frequency which is described in more detail in the following.

[0116] For the interpolation firstly an approximated si(x)-function is used which can be written as a simple polynom of order 2 according to the following approximation:

$$f(x) = \frac{\sin(x)}{x} \approx 1 - \frac{x^2}{6} \qquad (37)$$

[0117] Furthermore the autocorrelation coefficient is used for the interpolation at which the extended autocorrelation function $\hat{r}_{\tilde{y}\tilde{y},erw}(k,n)$ has the maximum, and also the adjacent autocorrelation coefficients unconsidered- that is the autocorrelation offsets left and right of the maximum. The interpolated speech fundamental period $\tau_{p,\mathrm{mod}}$ (n) can hereby be written as a function depending on the quantized speech fundamental period $\tau_p(n)$ and the considered autocorrelation coefficients according to the following equation:

$$\hat{\tau}_{\mathrm{p,mod}}(n) = \mathrm{Fkt}\Big(\tau_{\mathrm{p}}(n),\ \hat{r}_{\tilde{y}\tilde{y},\mathrm{erw}}\big(\tau_{\mathrm{p}}(n)-1,n\big),\ \hat{r}_{\tilde{y}\tilde{y},\mathrm{erw}}\big(\tau_{\mathrm{p}}(n),n\big),\ \hat{r}_{\tilde{y}\tilde{y},\mathrm{erw}}\big(\tau_{\mathrm{p}}(n)+1,n\big)\Big) \qquad (38)$$

[0118] In this context it has to be noted that if a correction according to the post-processing described in the previous section should be performed, the value $\tau_p(n)$ has to be replaced by the value $\tilde{\tau}_p(n)$. Finally the estimated and interpolated speech fundamental period can be determined according to

$$\hat{\tau}_{\mathrm{p,mod}}(n) = \tau_{\mathrm{p}}(n) \;-\; \Delta_p(n)\,, \qquad (39)$$

wherein $\Delta_p(n)$ is a correction value for the quantized speech fundamental period $\tau_p(n)$ which has to be determined in every frame clock n according to the following equation:

$$\Delta_p(n) = \frac{\hat{r}_{\tilde{y}\tilde{y},\mathrm{erw}}\big(\tau_\mathrm{p}(n)+1,n\big) - \hat{r}_{\tilde{y}\tilde{y},\mathrm{erw}}\big(\tau_\mathrm{p}(n)-1,n\big)}{2\Big(\hat{r}_{\tilde{y}\tilde{y},\mathrm{erw}}\big(\tau_\mathrm{p}(n)+1,n\big) + \hat{r}_{\tilde{y}\tilde{y},\mathrm{erw}}\big(\tau_\mathrm{p}(n)-1,n\big) - 2\,\hat{r}_{\tilde{y}\tilde{y},\mathrm{erw}}\big(\tau_\mathrm{p}(n),n\big)\Big)}. \qquad (40)$$

[0119] Finally the interpolation for improving the speech fundamental frequency estimation which is presented here shall be clarified in two examples. In the upper part of Fig. 16 the time-frequency-analysis of a portion of several sinusoidal signals of equal amplitude is shown. Contrary hereto a portion of a speech signal of a female voice is shown in the lower part of Fig. 16. The white graph denotes the estimated quantized speech fundamental frequency in the upper as well as also in the lower part of Fig. 16. The grey graph in the upper part respectively the black graph in the lower part demonstrates the estimated speech fundamental frequency after the interpolation. It can be seen from the upper part of Fig. 16 that due to the interpolation nearly the desired straight graph of the estimated speech fundamental frequency can be obtained. In the lower part it can be seen that the estimated speech fundamental frequency of the speech fundamental frequency structure follows the speech signal closely when the interpolation is used.

[0120] Furthermore the analysis has shown that an improvement of the speech fundamental frequency estimation of female voices up to about 30 Hz respectively with male voices about 10 Hz can be reached in the case the previously described interpolation is used.

[0121] Summarizing the problem presented in the introductory portion is solved presently in an approach having four independent steps each of which contributes to the total improvement and each of which can also be implemented independently from the others:

- For improvement of the spectral resolution short FIR-filters can be used in portions of the spectrum having low frequencies. This results in a significant improvement for medium speech fundamental frequencies.

- After the determination of necessary scaling values a noise reduction is performed. Thus, the method becomes more robust against background noise.

- In addition to the correlation of the actual signal frame with itself a correlation with the preceding signal frame is also calculated. However, significant error terms are generated hereby. By means of an adaptive correlation compensation those terms can be widely removed and the correlation mentioned second can thus be used for estimation of very low speech fundamental frequencies.

- By means of a simple interpolation a more precise estimation can be obtained. Finally erroneous detections which lead to doublings, respectively triplications, of the estimation are also corrected by means of adaptive post-processing.

[0122] Expressed in other words, this invention describes a method for estimating the fundamental frequency (pitch frequency) of speech signals. This is achieved in the DFT domain by analyzing the current input spectrum as well as past input spectra. To achieve an - compared to standard methods - improved estimation performance a four stage algorithm is applied or proposed whereby the steps can also be used independently: First, pre-processing (called spectral refinement) is applied to the input spectrum at low frequencies. Second, a noise reduction is applied when computing normalization values. Third, estimations for the autocorrelation of the current frame and cross correlation of the current with the previous frame are adaptively combined in order to obtain an extended range. Fourth, post-processing is applied to reduce estimation errors and to achieve an improved pitch accuracy.

Claims

1. Speech fundamental frequency estimator (1100) being configured for receiving a first set of values ($\tilde{Y}_1$) and a second set of values ($\tilde{Y}_2$), the first set of values ($\tilde{Y}_1$) being a frequency domain representation of a first set of time domain signal values (y_1) within a first time interval (t_1) and the second set of values ($\tilde{Y}_2$) being a frequency domain representation of a second set of time domain signal values (y_2) within a second time interval (t_2), the second time interval (t_2) being later than and offset from the first time interval (t_1), the speech fundamental frequency estimator (1100)

comprising:

- a first power density spectrum calculator (1102) being configured for storing a version of the first set of values ($\tilde{Y}_1$) and being configured for providing values of a first power density spectrum ($\hat{S}_{\tilde{y}\tilde{y}_d}(\Omega_\mu, n)$) by multiplying the stored version of the first set of values ($\tilde{Y}_1$) with a complex conjugate version of the second set of values ($\tilde{Y}_2$);
- a second power density spectrum calculator (1104) being configured for providing values of a second power density spectrum ($\hat{S}_{\tilde{y}\tilde{y}}(\Omega_\mu, n)$) by multiplying a version of the second set of values ($\tilde{Y}_2$) with a complex conjugate version of the the second set of values ($\tilde{Y}_2$);
- an analyzer (1106) being configured for determining the speech fundamental frequency estimate ($f_p(n)$) on the basis of the values of the first power density spectrum ($\hat{S}_{\tilde{y}\tilde{y}_d}(\Omega_\mu, n)$) and the values of the second power density spectrum ($\hat{S}_{\tilde{y}\tilde{y}}(\Omega_\mu, n)$).

2. Speech fundamental frequency estimator (1100) according to claim 1, **characterized in that** the first power density spectrum calculator (1102) is configured for multiplying versions of the sets of values ($\tilde{Y}_1$, $\tilde{Y}_2$) which represent sets of time domain signal values (y_1, y_2) having overlapping time intervals (t_1, t_2).

3. Speech fundamental frequency estimator (1100) according to claim 2, **characterized in that** the first power density spectrum calculator (1102) is configured for multiplying versions of the sets of values ($\tilde{Y}_1$, $\tilde{Y}_2$) which represent time domain signal values (y_1, y_2) having overlapping time intervals (t_1, t_2) of that least 25 percent.

4. Speech fundamental frequency estimator (1100) according to one of claims 1 to 3, **characterized in that** the second power density spectrum calculator (1104) is configured for providing a conjugate complex version of the second set of values ($\tilde{Y}_2$) to the first power density spectrum calculator (1102) and wherein the first power density spectrum calculator (1102) is configured for using the provided conjugate complex version of the second set of values ($\tilde{Y}_2$) as the version with which the stored a version of the first set of values ($\tilde{Y}_1$) is to be multiplied.

5. Speech fundamental frequency estimator (1100) according to any of the preceding claims, **characterized in that** the analyzer (1106) is configured for performing a first frequency-time-transform of the first power density spectrum ($\hat{S}_{\tilde{y}\tilde{y}_d}(\Omega_\mu, n)$) in order to obtain a first set of correlation function values ($\hat{r}_{\tilde{y}\tilde{y}_d,g}(m,n)$) and for performing a second frequency-time-transform of the second power density spectrum ($\hat{S}_{\tilde{y}\tilde{y}}(\Omega_\mu, n)$) in order to obtain a second set of correlation function values ($\hat{r}_{\tilde{y}\tilde{y},g}(m,n)$), wherein the analyzer (1106) is furthermore configured for determining a set of normalization values ($\bar{S}_{\tilde{y}\tilde{y}}(\Omega_\mu, n)$) and a set of weighting values *($V(e^{j\Omega_\mu}, n)$)* from the second power density spectrum ($\hat{S}_{\tilde{y}\tilde{y}}(\Omega_\mu, n)$ and for using the set of normalization values ($\bar{S}_{\tilde{y}\tilde{y}}(\Omega_\mu, n)$) and the set of weighting values *($V(e^{j\Omega_\mu}, n)$)* in the first and second frequency-time-transform and wherein the analyzer (1106) is furthermore configured for determining the speech fundamental frequency estimate ($f_p(n)$) on the basis of the first and second sets of correlation function values ($\hat{r}_{\tilde{y}\tilde{y}_d,g}(m,n)$, $\hat{r}_{\tilde{y}\tilde{y},g}(m,n)$).

6. Speech fundamental frequency estimator (1100) according to claim 5, **characterized in that** the analyzer (1106) further comprises a compensator being configured for adaptively compensating the values of the first set of correlation function values ($\hat{r}_{\tilde{y}\tilde{y}_d,g}(m,n)$) by a correction factor ($\Delta(m,n)$) being based on a value of the second set of

correlation function values $(\hat{r}_{\widetilde{yy},g}(m,n))$ and wherein the analyzer (1106) is furthermore configured for determining the speech fundamental frequency estimate ($f_p(n)$) on the basis of the compensated first set of correlation function values $(\hat{r}_{\widetilde{yy}_d,g,\mathrm{mod}}(m,n))$ and the second set of correlation function values $(\hat{r}_{\widetilde{yy},g}(m,n))$.

7. Speech fundamental frequency estimator (1100) according to claim 6, **characterized in that** the compensator is configured for multiplying the second set of correlation function values $(\hat{r}_{\widetilde{yy},g}(m,n))$ by a lower bounded quotient between a value of the first set of correlation function values $(\hat{r}_{\widetilde{yy}_d,g}(m,n))$ and a value of the second set of correlation function values $(\hat{r}_{\widetilde{yy},g}(m,n))$ in order to obtain said compensated first set of correlation function values $(\hat{r}_{\widetilde{yy}_d,g,\mathrm{mod}}(m,n))$.

8. Speech fundamental frequency estimator (1100) according to claim 7, **characterized in that** the analyzer (1106) is configured for combining the compensated first set of correlation function values $(\hat{r}_{\widetilde{yy}_d,g,\mathrm{mod}}(m,n))$ and the second set of correlation function values $(\hat{r}_{\widetilde{yy},g}(m,n))$ in order to obtain an extended set of correlation function values $(\hat{r}_{\widetilde{yy},erw}(k,n))$, wherein the values of the extended set of correlation function values $(\hat{r}_{\widetilde{yy},erw}(k,n))$ assume corresponding values from the compensated first set of correlation function values $(\hat{r}_{\widetilde{yy}_d,g,\mathrm{mod}}(m,n))$, the second set of correlation function values $(\hat{r}_{\widetilde{yy},g}(m,n))$ or values between the compensated first set of correlation function values $(\hat{r}_{\widetilde{yy}_d,g,\mathrm{mod}}(m,n))$ and the second set of correlation function values $(\hat{r}_{\widetilde{yy},g}(m,n))$ and wherein the analyzer (1106) is furthermore configured for determining the speech fundamental frequency estimate ($f_p(n)$) on the basis of said extended set of correlation function values $(\hat{r}_{\widetilde{yy},erw}(k,n))$.

9. Speech fundamental frequency estimator (1100) according to one of claims 5 to 8, **characterized in that** the analyzer (1106) is configured for determining the speech fundamental frequency estimate ($f_p(n)$) by searching the index of a maximum value (τ_p (n)) from the extended set of correlation function values $(\hat{r}_{\widetilde{yy},erw}(k,n))$ within a predetermined number of indices (k) of the values of the extended set of correlation values $(\hat{r}_{\widetilde{yy},erw}(k,n))$, from the first or second set of correlation function values $(\hat{r}_{\widetilde{yy}_d,g}(m,n),\ \hat{r}_{\widetilde{yy},g}(m,n))$ within a predetermined number of indexes (m) of values of the first respectively second set of correlation function values $(\hat{r}_{\widetilde{yy}_d,g}(m,n),\ \hat{r}_{\widetilde{yy},g}(m,n))$ or from the compensated first set of correlation function values $(\hat{r}_{\widetilde{yy}_d,g,\mathrm{mod}}(m,n))$ within the predetermined number of indices (m) of values of the compensated first set of correlation function values $(\hat{r}_{\widetilde{yy}_d,g,\mathrm{mod}}(m,n))$ and wherein the analyzer (1106) is furthermore configured for determining the speech fundamental frequency estimate ($f_p(n)$) as the product of a sampling frequency (f_s) and a reciprocal value of said searched index ($\tau_p(n)$).

10. Speech fundamental frequency estimator (1100) according to claim 9, **characterized in that** the analyzer (1106) is furthermore configured for determining a reliability factor (p_{f_p}(n)) for the determined speech fundamental frequency estimate and for blocking an output of the determined speech fundamental frequency estimate (f_p(n)) in the case the determined reliability factor ($p_{f_p}(n)$) for the determined speech fundamental frequency estimate is below a predetermined reliability factor (p_0).

11. Speech fundamental frequency estimator (1100) according to claim 10, **characterized in that** the analyzer (1106) is furthermore configured for determining said reliability factor ($p_{f_p}(n)$) by dividing the maximum value ($\hat{\tau}_p(n)$) at said searched by the first value of the extended set of correlation function values ($\hat{r}_{yy,erw}(k,n)$) or, respectively the first, the compensated first or second set of correlation function values

$(\hat{r}_{\widetilde{yy}_d,g}(m,n),\ \hat{r}_{\widetilde{yy}_d,g,\mathrm{mod}}(m,n),\ \hat{r}_{\widetilde{yy},g}(m,n))$.

12. Speech fundamental frequency estimator (1100) according to one of claims 5 to 11, **characterized in that** the second power density spectrum calculator (1104) is configured for determining an estimate of the power density spectrum of background noise ($\hat{S}_{nn}(\Omega_\mu, n)$) and for determining a noise suppression factor ($V(e^{j\Omega_\mu}, n)$) on the basis of said power density spectrum of background noise ($\hat{S}_{nn}(\Omega_\mu,n)$), and wherein the analyzer (1106) is configured for multiplying the first and second power density spectrum with said noise suppression factor ($V(e^{j\Omega_\mu}, n)$) prior to the frequency-time-transform of the first respectively second power density spectrum

$(\hat{S}_{\widetilde{yy}_d}(\Omega_\mu,n),\ \hat{S}_{\widetilde{yy}}(\Omega_\mu,n))$.

13. Speech fundamental frequency estimator (1100) according to claim 12, **characterized in that** the second power density spectrum calculator (1104) is configured for determining the noise suppression factor as the maximum of a predetermined maximum suppression coefficient (V_0) and a term being dependent on a quotient of the estimate of the power density spectrum of background noise ($\hat{S}_{nn}$ (Ω_μ,n)) and the second power density spectrum

$(\hat{S}_{\widetilde{yy}}(\Omega_\mu,n))$.

14. Speech fundamental frequency estimator (1100) according to one of claims 12 or 13, **characterized in that** the second power density spectrum calculator (1104) is configured for determining the estimate of the power density spectrum of background noise ($\hat{S}_{nn}(\Omega_\mu, n)$) in speech pauses or for determining the estimate of the power density spectrum of background noise ($\hat{S}_{nn}(\Omega_\mu, n)$) from a segment-wise estimation of the minima of the power of a differential signal.

15. Speech fundamental frequency estimator (1100) according to one of claims 12 to 14, **characterized in that** the noise suppression factor is defined by

$$V\left(e^{j\Omega_\mu}, n\right) = \max\left\{V_0,\ 1 - \beta\frac{\widehat{S}_{nn}(\Omega_\mu, n)}{\widehat{S}_{yy}(\Omega_\mu, n)}\right\}$$

wherein $\hat{S}_{nn}(\Omega_\mu, n)$ denotes the estimate of the power density spectrum of the background noise, $\hat{S}_{yy}(\Omega_\mu,n)$ denotes the second power density spectrum, V_0 denotes a predefined maximum attenuation factor and β denotes a value for overestimating the power density spectrum of the background noise ($\hat{S}_{nn}(\Omega_\mu,n)$).

16. Speech fundamental frequency estimator (1100) according to one of claims 5 to 15, **characterized in that** the analyzer (1106) is furthermore configured for reestimating the speech fundamental frequency estimate in the case the determined speech fundamental frequency estimate is below the predefined frequency value (f_k) wherein the analyzer (106) is configured for performing the reestimation by searching a further index (k, m) of a further maximum value ($\tilde{\tau}_p(n)$) of the extended set of correlation function values $(\hat{r}_{\widetilde{yy},erw}(k,n))$, the first or second set of correlation function values $(\hat{r}_{\tilde{y}\tilde{y}_d,g}(m,n), \hat{r}_{\tilde{y}\tilde{y},g}(m,n))$ or the compensated first set of correlation function values $(\hat{r}_{\tilde{y}\tilde{y}_d,g,\mathrm{mod}}(m,n))$ within a further number of values of said sets of correlation function values and for outputing a product of a sampling frequency (f_s) and a reciprocal value of said further index ($\hat{\tilde{\tau}}_p(n)$) as the determined speech fundamental frequency estimate.

17. Speech fundamental frequency estimator (1100) according to claim 16, **characterized in that** the analyzer (106) is configured for searching said index (k, m) of said further maximum value ($\tilde{\tau}_p(n)$) using a number of values k of said sets of correlation function values which is defined by

$$\frac{f_s}{f_{\mathrm{p,max}}} \leq k < \frac{f_s}{2f_{\mathrm{p}}(n)} + k_0$$

wherein k denotes the number of values of said sets of correlation function values, $f_p(n)$ denotes the previously determined speech fundamental frequency estimate, $f_{p,max}$ denotes a predefined value of a maximal possible speech fundamental frequency, f_s denotes a sampling frequency and k_0 denotes a constant.

18. Speech fundamental frequency estimator (1100) according to claim 16 or 17, **characterized in that** the analyzer (1106) is configured for outputting said product as the predetermined speech fundamental frequency estimate only in the case the further index ($\tilde{\tau}_p(n)$) is larger than 60 percent of the previously searched maximal index ($\tau_p(n)$) as well as a value $(\hat{r}_{\tilde{y}\tilde{y},erw}(\tilde{\tau}_p(n),n))$ of the extended set of correlation function values $(\hat{r}_{\tilde{y}\tilde{y},erw}(k,n))$ at said further index ($\tilde{\tau}_p(n)$) is larger than a previously defined amplitude value ($\tilde{p}_0$).

19. Speech fundamental frequency estimator (1100) according to one of claims 5 to 18, **characterized in that** the analyzer (1106) is configured for modifying a speech fundamental period ($\tilde{\tau}_p(n)$) corresponding to said determined speech fundamental frequency estimate by a interpolation correction term ($\Delta_p(n)$) prior of outputing a modified speech fundamental frequency estimate ($f_p(n)$), wherein said interpolation correction term (Δp) is dependent on values of said first or second set of correlation function values $(\hat{r}_{\tilde{y}\tilde{y}_d,g}(m,n), \hat{r}_{\tilde{y}\tilde{y},g}(m,n))$, of said extended set of correlation function values $(\hat{r}_{\tilde{y}\tilde{y},erw}(k,n))$ or said compensated first set of correlation function values $(\hat{r}_{\tilde{y}\tilde{y}_d,g,\mathrm{mod}}(m,n))$, respectively.

20. Speech fundamental frequency estimator (1100) according to one of claims 1 to 19, **characterized by** a frequency domain filtering unit being configured for receiving the frequency domain versions (Y_1, Y_2) of the first and second set of time domain signal values (y_1, Y_2), for frequency domain filtering said frequency domain versions in order to obtain said first and second sets of values ($\tilde{Y}_1, \tilde{Y}_2$), respectively, and for providing said first and second sets of values ($\tilde{Y}_1$, $\tilde{Y}_2$) to the first and second power density spectrum calculator respectively.

21. Speech fundamental frequency estimator (1100) according to claim 20, **characterized in that** the frequency domain filtering unit is configured for filtering only frequencies below a predefined limiting frequency.

22. Speech fundamental frequency estimator (1100) according to claim 21, **characterized in that** the frequency domain filtering unit is configured for delaying values of said frequency domain versions being above said predefined limiting frequency.

23. Method (1140) for estimating a speech fundamental frequency ($f_p(n)$), the method using a first set of values ($\tilde{Y}_1$)

and a second set of values ($\tilde{Y}_2$), the first set of values ($\tilde{Y}_1$) being a received frequency domain representation of a first set of time domain signal values (y_1) within a first time interval (t_1) and the second set of values ($\tilde{Y}_2$) being a received frequency domain representation of a second set of time domain signal values (y_2) within a second time interval (t_2), the second time interval (t_2) being later than and offset from the first time interval (t_1), the method for estimating the speech fundamental frequency ($f_p(n)$) comprising the steps of:

- storing (1150) a version of the first set of values ($\tilde{Y}_1$) and providing values of a first power density spectrum ($\hat{S}_{\widetilde{yy}_d}(\Omega_\mu, n)$) by multiplying (1152) the stored version of the first set of values ($\tilde{Y}_1$) with a complex conjugate version of the second set of values ($\tilde{Y}_2$);
- providing values of a second power density spectrum ($\hat{S}_{\widetilde{yy}}(\Omega_\mu, n)$) by multiplying (1153) a version of the second set of values ($\tilde{Y}_2$) with a complex conjugate version of the second set of values ($\tilde{Y}_2$);
- determining (1156) the speech fundamental frequency estimate (f_p) on the basis of the values of the first power density spectrum ($\hat{S}_{\widetilde{yy}_d}(\Omega_\mu, n)$) and the values of the second power density spectrum ($\hat{S}_{\widetilde{yy}}(\Omega_\mu, n)$).

24. Method (1140) according to claim 23, **characterized in that** the step of determining (1156) the speech fundamental frequency estimate ($f_p(n)$) comprises:

- • performing a first frequency-time-transform of the first power density spectrum ($\hat{S}_{\widetilde{yy}_d}(\Omega_\mu, n)$) in order to obtain a first set of correlation function values ($\hat{r}_{\widetilde{yy}_d,g}(m,n)$);
- • performing a second frequency-time-transform of the second power density spectrum ($\hat{S}_{\widetilde{yy}}(\Omega_\mu, n)$) in order to obtain a second set of correlation function values ($\hat{r}_{\widetilde{yy},g}(m,n)$), wherein the step of determining (1156) further comprises determining a set of normalization values ($\bar{S}_{\widetilde{yy}}(\Omega_\mu, n)$) and a set of weighting values ($V(e^{j\Omega_\mu}, n)$) from the second power density spectrum ($\hat{S}_{\widetilde{yy}}(\Omega_\mu, n)$ and using the set of normalization values ($\bar{S}_{\widetilde{yy}}(\Omega_\mu, n)$) ; and the set of weighting values ($V(e^{j\Omega_\mu}, n)$) in the first and second frequency-time-transform and wherein the determination of the speech fundamental frequency estimate ($f_p(n)$) is performed on the basis of the first and second sets of correlation function values ($\hat{r}_{\widetilde{yy}_d,g}(m,n), \hat{r}_{\widetilde{yy},g}(m,n)$).

25. Method (1140) according to claim 24, **characterized in that** the step of determining (1156) the speech fundamental frequency estimate ($f_p(n)$) comprises adaptively compensating the values of the first set of correlation function values ($\hat{r}_{\widetilde{yy}_d,g}(m,n)$) by a correction factor ($\Delta(m,n)$) being based on a value of the second set of correlation function values ($\hat{r}_{\widetilde{yy},g}(m,n)$) in order to obtain a compensated first set of values and determining the speech fundamental frequency estimate ($f_p(n)$) on the basis of the compensated first set of correlation function values ($\hat{r}_{\widetilde{yy}_d,g,\mathrm{mod}}(m,n)$) and the second set of correlation function values ($\hat{r}_{\widetilde{yy},g}(m,n)$).

26. Method (1140) according to claim 25, **characterized in that** the step of compensating comprises multiplying the second set of correlation function values ($\hat{r}_{\widetilde{yy},g}(m,n)$) by a lower bounded quotient between a value of the first set of correlation function values ($\hat{r}_{\widetilde{yy}_d,g}(m,n)$) and a value of the second set of correlation function values ($\hat{r}_{\widetilde{yy},g}(m,n)$) in order to obtain said compensated first set of correlation function values ($\hat{r}_{\widetilde{yy}_d,g,\mathrm{mod}}(m,n)$).

27. Method (1140) according to claim 26, **characterized in that** the step of determining (1156) the speech fundamental frequency estimate ($f_p(n)$) comprises combining the compensated first set of correlation function values ($\hat{r}_{\widetilde{yy}_d,g,\mathrm{mod}}(m,n)$) and the second set of correlation function values ($\hat{r}_{\widetilde{yy},g}(m,n)$) in order to obtain an extended set of correlation function values ($\hat{r}_{\widetilde{yy},erw}(k,n)$), wherein the values of the extended set of correlation function values ($\hat{r}_{\widetilde{yy},erw}(k,n)$) assume corresponding values from the compensated first set of correlation function values ($\hat{r}_{\widetilde{yy}_d,g,\mathrm{mod}}(m,n)$), the second set of correlation function values ($\hat{r}_{\widetilde{yy},g}(m,n)$) or values between the compensated first set of correlation function values ($\hat{r}_{\widetilde{yy}_d,g,\mathrm{mod}}(m,n)$) and the second set of correlation function values ($\hat{r}_{\widetilde{yy},g}(m,n)$) and wherein step of determining (1156) the speech fundamental frequency estimate ($f_p(n)$) further comprises determining the speech fundamental frequency estimate ($f_p(n)$) on the basis of said extended set of correlation function values ($\hat{r}_{\widetilde{yy},erw}(k,n)$).

28. Method (1140) according to one of claims 23 to 27, **characterized in that** the step of determining (1156) the speech fundamental frequency estimate ($f_p(n)$) comprises determining the speech fundamental frequency estimate ($f_p(n)$) by searching the index of a maximum value ($\tau_p(n)$) from the extended set of correlation function values ($\hat{r}_{\widetilde{yy},erw}(k,n)$) within a predetermined number of indices (k) of the values of the extended set of correlation values ($\hat{r}_{\widetilde{yy},erw}(k,n)$), from the first or second set of correlation function values ($\hat{r}_{\widetilde{yy}_d,g}(m,n)$, $\hat{r}_{\widetilde{yy},g}(m,n)$) within a predetermined number of indexes (m) of values of the first respectively second set of correlation function values ($\hat{r}_{\widetilde{yy}_d,g}(m,n)$, $\hat{r}_{\widetilde{yy},g}(m,n)$) or from the compensated first set of correlation function values ($\hat{r}_{\widetilde{yy}_d,g,\mathrm{mod}}(m,n)$) within the predetermined number of indices (m) of values of the compensated first set of correlation function values ($\hat{r}_{\widetilde{yy}_d,g,\mathrm{mod}}(m,n)$) and wherein the step of determining (1156) the speech fundamental frequency estimate ($f_p(n)$) furthermore comprises determining the speech fundamental frequency estimate ($f_p(n)$) as the product of a sampling frequency (f_s) and a reciprocal value of said searched index *(τ_p (n)).*

29. Method (1140) according to claim 28, **characterized in that** the step of determining (1156) the speech fundamental frequency estimate ($f_p(n)$) comprises determining a reliability factor ($p_{f_p}(n)$) for the determined speech fundamental frequency estimate ($f_p(n)$) and for blocking an output of the determined speech fundamental frequency estimate ($f_p(n)$) in the case the determined reliability factor (p_{f_p} (n)) for the determined speech fundamental frequency estimate ($f_p(n)$) is below predetermined reliability factor (p_0).

30. Method (1140) according to claim 29, **characterized in that** the step of determining (1156) the speech fundamental frequency estimate ($f_p(n)$) comprises determining said reliability factor ($p_{f_p}(n)$) by dividing the maximum value ($\tilde{\tau}_p$

(n)) at said searched by the first value of the extended set of correlation function values $(\hat{r}_{\tilde{y}\tilde{y},erw}(k,n))$ or, respectively the first, the compensated first or second set of correlation function values $(\hat{r}_{\tilde{y}\tilde{y}_d,g}(m,n),\ \hat{r}_{\tilde{y}\tilde{y}_d,g,\mathrm{mod}}(m,n),\ \hat{r}_{\tilde{y}\tilde{y},g}(m,n))$.

31. Method (1140) according to one of claims 23 to 30, **characterized in that** the step of providing values of a second power density spectrum $(\hat{S}_{\tilde{y}\tilde{y}}(\Omega_\mu,n))$ comprises determining an estimate of the power density spectrum of background noise ($\hat{S}_{nn}(\Omega_\mu,\ n)$) and determining a noise suppression factor ($V(e^{j\Omega_\mu},n)$) on the basis of said power density spectrum of background noise ($\hat{S}_{nn}(\Omega_\mu,\ n)$), and the step of determining (1156) the speech fundamental frequency estimate (f_p(n)) comprises multiplying the first and second power density spectrum with said noise suppression factor ($V(e^{j\Omega_\mu},\ n)$) prior to the frequency-time-transform of the first respectively second power density spectrum $(\hat{S}_{\tilde{y}\tilde{y}_d}(\Omega_\mu,n),\ \hat{S}_{\tilde{y}\tilde{y}}(\Omega_\mu,n))$.

32. Method (1140) according to claim 31, **characterized in that** the step of providing values of a second power density spectrum $(\hat{S}_{\tilde{y}\tilde{y}}(\Omega_\mu,n))$ comprises determining the noise suppression factor as the maximum of a predetermined maximum suppression coefficient (V_0) and a term being dependent on a quotient of the estimate of the power density spectrum of background noise ($\hat{S}_{nn}(\Omega_\mu,\ n)$) and the second power density spectrum $(\hat{S}_{\tilde{y}\tilde{y}}(\Omega_\mu,n))$.

33. Method (1140) according to claim 32, **characterized in that** the step of providing values of a second power density spectrum $(\hat{S}_{\tilde{y}\tilde{y}}(\Omega_\mu,n))$ comprises determining the estimate of the power density spectrum of background noise ($\hat{S}_{nn}\ (\Omega_\mu,\ n)$) in speech pauses or for determining the estimate of the power density spectrum of background noise ($\hat{S}_{nn}\ (\Omega_\mu,\ n)$) from a segment-wise estimation of the minima of the power of a differential signal.

34. Method (1140) according to one of claims 31 to 33, **characterized in that** the noise suppression factor is defined by

$$V\left(e^{j\Omega_\mu},n\right) = \max\left\{V_0,\ 1-\beta\frac{\widehat{S}_{nn}(\Omega_\mu,n)}{\widehat{S}_{yy}(\Omega_\mu,n)}\right\}$$

wherein $\hat{S}_{nn}(\Omega_\mu,\ n)$ denotes the estimate of the power density spectrum of the background noise, $\hat{S}_{yy}(\Omega_\mu,n)$ denotes the second power density spectrum, V_0 denotes a predefined maximum attenuation factor and β denotes a value for overestimating the power density spectrum of the background noise ($\hat{S}_{nn}(\Omega_\mu,\ n)$).

35. Method (1140) according to one of claims 24 to 34, **characterized in that** the step of determining (1156) the speech fundamental frequency estimate (f_p(n)) comprises reestimating the speech fundamental frequency estimate (f_p(n)) in the case the determined speech fundamental frequency estimate is below the predefined frequency value (f_k) wherein the step of determining (1156) the speech fundamental frequency estimate (f_p(n)) comprises performing the reestimation by searching a further index (k, m) of a further maximum value ($\tilde{\tau}_p(n)$) of the extended set of correlation function values $(\hat{r}_{\tilde{y}\tilde{y},erw}(k,n))$, the first or second set of correlation function values $(\hat{r}_{\tilde{y}\tilde{y}_d,g}(m,n),\hat{r}_{\tilde{y}\tilde{y},g}(m,n))$ or the compensated first set of correlation function values $(\hat{r}_{\tilde{y}\tilde{y}_d,g,\mathrm{mod}}(m,n))$ within a further number of values of said sets of correlation function values and outputing a product of a sampling frequency (f_s) and a reciprocal value of said further index ($\tilde{\tau}_p$(n)) as the determined speech fundamental frequency estimate.

36. Method (1140) according to claim 35, **characterized in that** the step of determining (1156) the speech fundamental frequency estimate (f_p(n)) comprises searching said index (k, m) of said further maximum value ($\tilde{\tau}_p(n)$) using a number of values k of said sets of correlation function values which is defined by

$$\frac{f_s}{f_{\mathrm{p,max}}} \leq k < \frac{f_s}{2f_{\mathrm{p}}(n)} + k_0$$

wherein k denotes the number of values of said sets of correlation function values, f_p(n) denotes the previously determined speech fundamental frequency estimate, $f_{p,max}$ denotes a predefined value of a maximal possible speech fundamental frequency, f_s denotes a sampling frequency and k_0 denotes a constant.

37. Method (1140) according to one of claims 35 or 36, **characterized in that** the step of determining (1156) the speech fundamental frequency estimate (f_p(n)) comprises outputing said product as the predetermined speech fundamental frequency estimate (f_p(n)) only in the case that the autocorrelation function at the further index ($\tilde{\tau}_p$(n)) is larger than 60 percent of the autocorrelation function at the previously searched maximal index ($\tau_p(n)$) as well as the value $(\hat{r}_{\tilde{y}\tilde{y},erw}(\tilde{\tau}_p(n),n))$ of the extended set of correlation function values $(\hat{r}_{\tilde{y}\tilde{y},erw}(k,n))$ at said further index ($\tilde{\tau}_p$ (n)) is larger than a previously defined amplitude value ($\tilde{p}_0$).

38. Method (1140) according to one of claims 24 or 37, **characterized in that** the step of determining the speech fundamental frequency estimate (f_p(n)) comprises modifying a speech fundamental period ($\tilde{\tau}_p(n)$, $\tilde{\tau}_p(n)$) corresponding to said determined speech fundamental frequency estimate (f_p(n)) by a interpolation correction term (Δ_p(n)) prior of outputing said speech fundamental frequency estimate (f_p(n)), wherein said interpolation correction term (Δ_p(n)) is dependent on values of said first or second set of correlation function values $(\hat{r}_{\tilde{y}\tilde{y}_d,g}(m,n), \hat{r}_{\tilde{y}\tilde{y},g}(m,n))$, of said extended set of correlation function values $(\hat{r}_{\tilde{y}\tilde{y},erw}(k,n))$ or said compensated first set of correlation function values $(\hat{r}_{\tilde{y}\tilde{y}_d,g,\mathrm{mod}}(m,n))$, respectively.

39. Method (1140) according to one of the preceding claims, **characterized in that** the method further comprises a step of receiving the frequency domain versions (Y_1, Y_2) of the first and second set of time domain signal values (y_1, y_2), frequency domain filtering said frequency domain versions in order to obtain said first and second sets of values ($\tilde{Y}_1$, $\tilde{Y}_2$), respectively, and providing said first and second sets of values ($\tilde{Y}_1$, $\tilde{Y}_2$) the first and second power density spectrum calculator respectively.

40. Method (1140) according to claim 39, **characterized in that** the step of frequency domain filtering is only performed for frequencies below a predefined limiting frequency.

41. Method (1140) according to claim 40, **characterized in that** the step of frequency domain filtering comprises delaying values of said frequency domain versions being above said predefined limiting frequency.

42. Computer program having a program code for performing the method according to one of claims 23 to 41, when the computer program runs on a computer. '

43. Speech fundamental frequency estimator (900), being configured for receiving a set of values ($\tilde{Y}_1$), the set of values

($\tilde{Y}_1$) being a frequency domain representation of a set of time domain signal values ($\tilde{Y}_1$) within a time interval (t_1), the speech fundamental frequency estimator (900) comprising:

- a power density spectrum calculator (902) being configured for providing values of a power density spectrum ($\hat{S}_{\tilde{y}\tilde{y}}(\Omega_\mu, n)$) by multiplying a version of the set of values ($\tilde{Y}_2$) with a compley conjugate version of the set of values ($\tilde{Y}_2$), wherein the power density spectrum calculator (902) is configured for determining an estimate of the power density spectrum of background noise ($\hat{S}_{nn}(\Omega_\mu,n)$) and for determining a noise suppression factor ($V(e^{j\Omega_\mu},n)$) on the basis of said power density spectrum of background noise ($\hat{S}_{nn}(\Omega_\mu, n)$);

- an analyzer (904) being configured multiplying the power density spectrum ($\hat{S}_{\tilde{y}\tilde{y}}(\Omega_\mu, n)$) with said noise suppression factor ($V(e^{j\Omega_\mu},n)$) and for performing a frequency-time-transform of the multiplied values of the power density spectrum ($\hat{S}_{\tilde{y}\tilde{y}}(\Omega_\mu, n)$) in order to obtain a set of correlation function values ($\hat{r}_{\tilde{y}\tilde{y}}(k, n)$), wherein the analyzer (904) is furthermore configured for determining the speech fundamental frequency estimate ($f_p(n)$) on the basis of the set of correlation function values ($\hat{r}_{\tilde{y}\tilde{y}}(k, n)$).

44. Speech fundamental frequency estimator (900) according to claim 43, **characterized in that** the power density spectrum calculator (902) is configured for determining the noise suppression factor as the maximum of a predetermined maximum suppression coefficient (V_0) and a term being dependent on a quotient of the estimate of the power density spectrum of background noise ($\hat{S}_{nn}(\Omega_\mu, n)$) and the second power density spectrum ($\hat{S}_{\tilde{y}\tilde{y}}(\Omega_\mu, n)$).

45. Speech fundamental frequency estimator (900) according to one of claims 43 or 44, **characterized in that** the power density spectrum calculator (902) is configured for determining the estimate of the power density spectrum of background noise ($\hat{S}_{nn}(\Omega_\mu, n)$) in speech pauses or for determining the estimate of the power density spectrum of background noise ($\hat{S}_{nn}(\Omega_\mu, n)$) from a segment-wise estimation of the minima of the power of the differential signal.

46. Speech fundamental frequency estimator (900) according to one of claims 43 to 45, **characterized in that** the noise suppression factor is defined by

$$V\left(e^{j\Omega_\mu}, n\right) = \max\left\{V_0,\ 1 - \beta\frac{\widehat{S}_{nn}(\Omega_\mu, n)}{\widehat{S}_{yy}(\Omega_\mu, n)}\right\}$$

wherein $\hat{S}_{nn}(\Omega_\mu, n)$ denotes the estimate of the power density spectrum of the background noise, $\hat{S}_{yy}(\Omega_\mu, n)$ denotes the second power density spectrum, V_0 denotes a predefined maximum attenuation factor and β denotes a value for overestimating the power density spectrum of the background noise $(\hat{S}_{nn}(\Omega_\mu,n))$.

47. Method (940) for estimating a speech fundamental frequency ($f_p(n)$), the method being configured for receiving a set of values ($\tilde{Y}_1$), the set of values ($\tilde{Y}_1$) being a frequency domain representation of a set of time domain signal values (y_1) within a time interval (t_1), the method comprising the steps of:

- providing (950) values of a power density spectrum ($\hat{S}_{\tilde{y}\tilde{y}}(\Omega_\mu, n)$) by multiplying a version of the set of values ($\tilde{Y}_2$) with a a complex conjugate version of the set of values ($\tilde{Y}_2$),
- determining (952) an estimate of the power density spectrum of background noise ($\hat{S}_{nn}(\Omega_\mu,n)$) and determining (954) a noise suppression factor ($V(e^{j\Omega}{}_\mu$,n)) on the basis of said power density spectrum of background noise $(\hat{S}_{nn}(\Omega_\mu, n))$;

- multiplying (956) the power density spectrum $(\hat{S}_{\tilde{y}\tilde{y}}(\Omega_\mu, n))$ with said noise suppression factor $(V(e^{j\Omega_\mu}, n))$;

- performing (958) a frequency-time-transform of the multiplied values of the power density spectrum $(\hat{S}_{\tilde{y}\tilde{y}}(\Omega_\mu, n))$ in order to obtain a set of correlation function values $(\hat{r}_{\tilde{y}\tilde{y}}(k, n))$ and

- determining (960) the speech fundamental frequency estimate (f_p) on the basis of the set of correlation function values $(\hat{r}_{\tilde{y}\tilde{y}}(k, n))$.

48. Method (940) for estimating a speech fundamental frequency according to claim 47, **characterized in that** the step of determining (952) a noise suppression factor $(V(e^{j\Omega_\mu}, n))$ comprises determining the noise suppression factor as the maximum of a predetermined maximum suppression coefficient (V_0) and a term being dependent on a quotient of the estimate of the power density spectrum of background noise $(\hat{S}_{nn}(\Omega_\mu, n))$ and the second power density spectrum $(\hat{S}_{\tilde{y}\tilde{y}}(\Omega_\mu, n))$.

49. Method (940) according to claim 48, **characterized in that** the step of determining (952) a noise suppression factor $(V(e^{j\Omega_\mu}, n))$ comprises determining the estimate of the power density spectrum of background noise $(\hat{S}_{nn}(\Omega_\mu, n))$ in speech pauses or determining the estimate of the power density spectrum of background noise $(\hat{S}_{nn}(\Omega_\mu, n))$ from a segment-wise estimation of the minima of the power of the differential signal.

50. Method (940) according to claim 49, **characterized in that** the step of determining (952) noise suppression factor $(V(e^{j\Omega_\mu}, n))$ is **characterized in that** the noise suppression factor is defined by

$$V\left(e^{j\Omega_\mu}, n\right) = \max\left\{V_0,\ 1 - \beta\frac{\widehat{S}_{nn}(\Omega_\mu, n)}{\widehat{S}_{yy}(\Omega_\mu, n)}\right\}$$

wherein $\hat{S}_{nn}(\Omega_\mu, n)$ denotes the estimate of the power density spectrum of the background noise, $\hat{S}_{yy}(\Omega_\mu, n)$ denotes the second power density spectrum, V_0 denotes a predefined maximum attenuation factor and β denotes a value for overestimating the power density spectrum of the background noise $(\hat{S}_{nn}(\Omega_\mu, n))$.

51. Computer program having a program code for performing the method according to one of claims 47 to 50, when the computer program runs on a computer.

Amended claims in accordance with Rule 137(2) EPC.

1. Speech fundamental frequency estimator (1100) being configured for receiving a first set of values ($\tilde{Y}_1$) and a second set of values ($\tilde{Y}_2$), the first set of values ($\tilde{Y}_1$) being a frequency domain representation of a first set of time domain signal values (y_1) within a first time interval (t_1) and the second set of values ($\tilde{Y}_2$) being a frequency domain representation of a second set of time domain signal values (y_2) within a second time interval (t_2), the second time interval (t_2) being later than and offset from the first time interval (t_1), the speech fundamental frequency estimator (1100) comprising:

- a first power density spectrum calculator (1102) being configured for storing a version of the first set of values ($\tilde{Y}_1$) and being configured for providing values of a first power density spectrum $(\hat{S}_{yyd}(\Omega_\mu, n))$ by multiplying the stored version of the first set of values ($\tilde{Y}_1$) with a complex conjugate version of the second set of values ($\tilde{Y}_2$);

- a second power density spectrum calculator (1104) being configured for providing values of a second power density spectrum ($\hat{S}_{yy}(\Omega_\mu, n)$) by multiplying a version of the second set of values ($\tilde{Y}_2$) with a complex conjugate version of the the second set of values ($\tilde{Y}_2$);
- an analyzer (1106) being configured for determining the speech fundamental frequency estimate ($f_p(n)$) on the basis of the values of the first power density spectrum ($\hat{S}_{yyd}(\Omega_\mu, n)$) and the values of the second power density spectrum ($\hat{S}_{yy}(\Omega_\mu, n)$),

wherein the analyzer is further configured
for performing a first frequency-time-transform of the first power density spectrum ($\hat{S}_{yyd}(\Omega_\mu, n)$) in order to obtain a first set of correlation function values ($\hat{r}_{yyd,g}(m, n)$),
for performing a second frequency-time-transform of the second power density spectrum ($\hat{S}_{yy}(\Omega_\mu, n)$) in order to obtain a second set of correlation function values ($\hat{r}_{yy,g}(m,n)$), and
for determining the speech fundamental frequency estimate ($f_p(n)$) on the basis of the first and second sets of correlation function values ($\hat{r}_{yyd,g}(m,n)$, $\hat{r}_{yy,g}(m,n)$).

2. Speech fundamental frequency estimator (1100) according to claim 1, **characterized in that** the first power density spectrum calculator (1102) is configured for multiplying versions of the sets of values ($\tilde{Y}_1$, $\tilde{Y}_2$) which represent sets of time domain signal values (y_1, y_2) having overlapping time intervals (t_1, t_2).

3. Speech fundamental frequency estimator (1100) according to claim 2, **characterized in that** the first power density spectrum calculator (1102) is configured for multiplying versions of the sets of values ($\tilde{Y}_1$, $\tilde{Y}_2$) which represent time domain signal values (y_1, y_2) having overlapping time intervals (t_1, t_2) of that least 25 percent.

4. Speech fundamental frequency estimator (1100) according to one of claims 1 to 3, **characterized in that** the second power density spectrum calculator (1104) is configured for providing a conjugate complex version of the second set of values ($\tilde{Y}_2$) to the first power density spectrum calculator (1102) and
wherein the first power density spectrum calculator (1102) is configured for using the provided conjugate complex version of the second set of values ($\tilde{Y}_2$) as the version with which the stored a version of the first set of values ($\tilde{Y}_1$) is to be multiplied.

5. Speech fundamental frequency estimator (1100) according to any of the preceding claims, **characterized in that** the analyzer (1106) is configured for performing a first frequency-time-transform of the first power density spectrum ($\hat{S}_{yyd}(\Omega_\mu, n)$) in order to obtain a first set of correlation function values ($\hat{r}_{yyd,g}(m, n)$) and for performing a second frequency-time-transform of the second power density spectrum ($\hat{S}_{yy}(\Omega_\mu, n)$) in order to obtain a second set of correlation function values ($\hat{r}_{yy,g}(m,n)$), wherein the analyzer (1106) is furthermore configured for determining a set of normalization values ($\hat{S}_{yy}(\Omega_\mu, n)$) and a set of weighting values ($V(e^{j\Omega_\mu}, n)$) from the second power density spectrum ($\hat{S}_{yy}(\Omega_\mu, n)$ and for using the set of normalization values ($\hat{S}_{yy}(\Omega_\mu, n)$) and the set of weighting values ($V(e^{j\Omega_\mu}, n)$) in the first and second frequency-time-transform and wherein the analyzer (1106) is furthermore configured for determining the speech fundamental frequency estimate ($f_p(n)$) on the basis of the first and second sets of correlation function values ($\hat{r}_{yyd,g}(m, n)$, $\hat{r}_{yy,g}(m, n)$).

6. Speech fundamental frequency estimator (1100) according to claim 5, **characterized in that** the analyzer (1106) further comprises a compensator being configured for adaptively compensating the values of the first set of correlation function values ($\hat{r}_{yyd,g}(m, n)$) by a correction factor ($\Delta(m,n)$) being based on a value of the second set of correlation function values ($\hat{r}_{yy,g}(m,n)$) and wherein the analyzer (1106) is furthermore configured for determining the speech fundamental frequency estimate ($f_p(n)$) on the basis of the compensated first set of correlation function values ($\hat{r}_{yyd,g,mod}(m, n)$) and the second set of correlation function values ($\hat{r}_{yy,g}(m,n)$).

7. Speech fundamental frequency estimator (1100) according to claim 6, **characterized in that** the compensator is configured for multiplying the second set of correlation function values ($\hat{r}_{yy,g}(m, n)$) by a lower bounded quotient between a value of the first set of correlation function values ($\hat{r}_{yyd,g}(m, n)$) and a value of the second set of correlation function values ($\hat{r}_{yy,g}(m,n)$) in order to obtain said compensated first set of correlation function values ($\hat{r}_{yyd,g,mod}(m, n)$).

8. Speech fundamental frequency estimator (1100) according to claim 7, **characterized in that** the analyzer (1106) is configured for combining the compensated first set of correlation function values ($\hat{r}_{yyd,g,mod}(m,n)$) and the second set of correlation function values ($\hat{r}_{yy,g}(m, n)$) in order to obtain an extended set of correlation function values ($\hat{r}_{yy,erw}$

(k,n)), wherein the values of the extended set of correlation function values ($\hat{r}_{yy,erw}$ (k, n)) assume corresponding values from the compensated first set of correlation function values ($\hat{r}_{yyd,g,mod}$ (m,n)), the second set of correlation function values ($\hat{r}_{yy,g}$ (m, n)) or values between the compensated first set of correlation function values ($\hat{r}_{yyd,g,mod}$ (m,n)) and the second set of correlation function values ($\hat{r}_{yy,g}$ (m, n)) and wherein the analyzer (1106) is furthermore configured for determining the speech fundamental frequency estimate (f_p(n)) on the basis of said extended set of correlation function values ($\hat{r}_{yy,erw}(k,n)$).

9. Speech fundamental frequency estimator (1100) according to one of claims 5 to 8, **characterized in that** the analyzer (1106) is configured for determining the speech fundamental frequency estimate (f_p(n)) by searching the index of a maximum value ($\tau_p(n)$) from the extended set of correlation function values ($\hat{r}_{yy}$,*erw*,(k,n)) within a predetermined number of indices (k) of the values of the extended set of correlation values ($\hat{r}_{yy,erw}(k,n)$), from the first or second set of correlation function values ($\hat{r}_{yyd,g}$ (m, n), $\hat{r}_{yy,g}(m,n)$) within a predetermined number of indexes (m) of values of the first respectively second set of correlation function values ($\hat{r}_{yyd,g}$ (m, n), $\hat{r}_{yy,g}$ (m,n)) or from the compensated first set of correlation function values ($\hat{r}_{yyd,g,\mathrm{mod}}$ (m, n)) within the predetermined number of indices (m) of values of the compensated first set of correlation function values ($\hat{r}_{yyd,g,\ mod}$ (m, n)) and wherein the analyzer (1106) is furthermore configured for determining the speech fundamental frequency estimate (f_p(n)) as the product of a sampling frequency (f_s) and a reciprocal value of said searched index ($\tau_p(n)$).

10. Speech fundamental frequency estimator (1100) according to claim 9, **characterized in that** the analyzer (1106) is furthermore configured for determining a reliability factor ($p_{f_p}(n)$) for the determined speech fundamental frequency estimate and for blocking an output of the determined speech fundamental frequency estimate (f_p(n)) in the case the determined reliability factor ($p_{f_p}(n)$) for the determined speech fundamental frequency estimate is below a predetermined reliability factor (p_0).

11. Speech fundamental frequency estimator (1100) according to claim 10, **characterized in that** the analyzer (1106) is furthermore configured for determining said reliability factor (p_{f_p} (n)) by dividing the maximum value ($\tilde{\tau}_p$ (n)) at said searched index by the first value of the extended set of correlation function values ($\hat{r}_{yy,erw}$ (k,n)) or, respectively the first, the compensated first or second set of correlation function values ($\hat{r}_{yyd,g}$ (m, n), $\hat{r}_{yy,g,mod}$ (m, n), $\hat{r}_{yy,g}$ (m,n)).

12. Speech fundamental frequency estimator (1100) according to one of claims 5 to 11, **characterized in that** the second power density spectrum calculator (1104) is configured for determining an estimate of the power density spectrum of background noise ($\hat{S}_{nn}(\Omega_\mu,n)$) and for determining a noise suppression factor ($V(e^{j\Omega_\mu},n)$) on the basis of said power density spectrum of background noise ($S_{nn}(\Omega_\mu,n)$), and wherein the analyzer (1106) is configured for multiplying the first and second power density spectrum with said noise suppression factor ($V(e^{j\Omega_\mu},n)$) prior to the frequency-time-transform of the first respectively second power density spectrum ($\hat{S}_{yyd}(\Omega_\mu,n)$, $\hat{S}_{yy}(\Omega_\mu, n)$).

13. Speech fundamental frequency estimator (1100) according to claim 12, **characterized in that** the second power density spectrum calculator (1104) is configured for determining the noise suppression factor as the maximum of a predetermined maximum suppression coefficient (Vo) and a term being dependent on a quotient of the estimate of the power density spectrum of background noise ($\hat{S}_{nn}$ (Ω_μ,n)) and the second power density spectrum ($\hat{S}_{yy}$ (Ω_μ,n)).

14. Speech fundamental frequency estimator (1100) according to one of claims 12 or 13, **characterized in that** the second power density spectrum calculator (1104) is configured for determining the estimate of the power density spectrum of background noise ($\hat{S}_{nn}$ (Ω_μ, n)) in speech pauses or for determining the estimate of the power density spectrum of background noise (S_{nn} (Ω_μ,n)) from a segment-wise estimation of the minima of the power of a differential signal.

15. Speech fundamental frequency estimator (1100) according to claim 13 or claims 13 and 14, **characterized in that** the noise suppression factor is defined by

$$V\left(e^{j\Omega_\mu}, n\right) = \max\left\{V_0,\ 1-\beta\frac{\widehat{S}_{nn}(\Omega_\mu,n)}{\widehat{S}_{yy}(\Omega_\mu,n)}\right\}$$

wherein $\hat{S}_{nn}(\Omega_\mu,n)$ denotes the estimate of the power density spectrum of the background noise, $\hat{S}_{yy}(\Omega_\mu,n)$ denotes the second power density spectrum, V_0 denotes a predefined maximum attenuation factor and β denotes a value for overestimating the power density spectrum of the background noise ($\hat{S}_{nn}(\Omega_\mu, n)$).

16. Speech fundamental frequency estimator (1100) according to one of claims 5 to 15, **characterized in that** the analyzer (1106) is furthermore configured for reestimating the speech fundamental frequency estimate in the case the determined speech fundamental frequency estimate is below the predefined frequency value (f_k) wherein the analyzer (106) is configured for performing the reestimation by searching a further index (k, m) of a further maximum value ($\hat{\tau}_p(n)$) of the extended set of correlation function values ($\hat{r}_{yy,erw}(k,n)$), the first or second set of correlation function values ($\hat{r}_{yyd,g}$ (m, n), $\hat{r}_{yy,g}$ (m, n)) or the compensated first set of correlation function values ($\hat{r}_{yyd,g,mod}$ (m, n)) within a further number of values of said sets of correlation function values and for outputing a product of a sampling frequency (f_s) and a reciprocal value of said further index ($\tilde{\tau}_p(n)$) as the determined speech fundamental frequency estimate.

17. Speech fundamental frequency estimator (1100) according to claim 16, **characterized in that** the analyzer (106) is configured for searching said index (k, m) of said further maximum value ($\tilde{\tau}_p(n)$ using a number of values k of said sets of correlation function values which is defined by

$$\frac{f_s}{f_{\mathrm{p,max}}} \leq k < \frac{f_s}{2f_{\mathrm{p}}(n)} + k_0$$

wherein k denotes the number of values of said sets of correlation function values, $f_p(n)$ denotes the previously determined speech fundamental frequency estimate, $f_{p,max}$ denotes a predefined value of a maximal possible speech fundamental frequency, f_s denotes a sampling frequency and k_0 denotes a constant.

18. Speech fundamental frequency estimator (1100) according to claim 16 or 17, **characterized in that** the analyzer (1106) is configured for outputting said product as the predetermined speech fundamental frequency estimate only in the case the further index ($\tilde{\tau}_p(n)$) is larger than 60 percent of the previously searched maximal index ($\tau_p(n)$) as well as a value ($\hat{r}_{yy,erw}(\tilde{\tau}_p(n),n)$) of the extended set of correlation function values ($\hat{r}_{yy,erw}$ (k,n)) at said further index ($\tilde{\tau}_p(n)$) is larger than a previously defined amplitude value ($\tilde{p}_0$).

19. Speech fundamental frequency estimator (1100) according to one of claims 5 to 18, **characterized in that** the analyzer (1106) is configured for modifying a speech fundamental period ($\hat{\tau}_p(n)$) corresponding to said determined speech fundamental frequency estimate by a interpolation correction term ($\Delta_p(n)$) prior of outputing a modified speech fundamental frequency estimate (fp(n)), wherein said interpolation correction term (Δ_p) is dependent on values of said first or second set of correlation function values ($\hat{r}_{yyd,g}(m,n)$, $\hat{r}_{yy,g}(m,n)$), of said extended set of correlation function values ($\hat{r}_{yy,erw}(k,n)$) or said compensated first set of correlation function values ($\hat{r}_{yyd,g,mod}$ (m, n)), respectively.

20. Speech fundamental frequency estimator (1100) according to one of claims 1 to 19, **characterized by** a frequency domain filtering unit being configured for receiving the frequency domain versions (Y_1, Y_2) of the first and second set of time domain signal values (y_1, y_2), for frequency domain filtering said frequency domain versions in order to obtain said first and second sets of values ($\tilde{Y}_1, \tilde{Y}_2$), respectively, and for providing said first and second sets of values ($\tilde{Y}_1$, $\tilde{Y}_2$) to the first and second power density spectrum calculator respectively.

21. Speech fundamental frequency estimator (1100) according to claim 20, **characterized in that** the frequency domain filtering unit is configured for filtering only frequencies below a predefined limiting frequency.

22. Speech fundamental frequency estimator (1100) according to claim 21, **characterized in that** the frequency domain filtering unit is configured for delaying values of said frequency domain versions being above said predefined limiting frequency.

23. Method (1140) for estimating a speech fundamental frequency ($f_p(n)$), the method using a first set of values ($\tilde{Y}_1$) and a second set of values ($\tilde{Y}_2$), the first set of values ($\tilde{Y}_1$) being a received frequency domain representation of a first set of time domain signal values (y_1) within a first time interval (t_1) and the second set of values ($\tilde{Y}_2$) being a received frequency domain representation of a second set of time domain signal values (y_2) within a second time interval (t_2), the second time interval (t_2) being later than and offset from the first time interval (t_1), the method for estimating the speech fundamental frequency ($f_p(n)$) comprising the steps of:

- storing (1150) a version of the first set of values ($\tilde{Y}_1$) and providing values of a first power density spectrum ($\hat{S}_{yy_d}$ (Ω_μ, n)) by multiplying (1152) the stored version of the first set of values ($\tilde{Y}_1$) with a complex conjugate version of the second set of values ($\tilde{Y}_2$);
- providing values of a second power density spectrum ($\hat{S}_{yy}(\Omega_\mu, n)$) by multiplying (1153) a version of the second set of values ($\tilde{Y}_2$) with a complex conjugate version of the second set of values ($\tilde{Y}_2$);
- determining (1156) the speech fundamental frequency estimate (f_p) on the basis of the values of the first power density spectrum ($\hat{S}_{yy_d}(\Omega_\mu,n)$) and the values of the second power density spectrum ($\hat{S}_{yy}(\Omega_\mu,n)$),

wherein the step of determining the speech fundamental frequency estimate ($f_p(n)$) comprises

performing a first frequency-time-transform of the first power density spectrum ($\hat{S}_{yy}(\Omega_\mu$ $n)$) in order to obtain a first set of correlation function values ($\hat{r}_{yy_d,g}$ (m,n)),
performing a second frequency-time-transform of the second power density spectrum ($\hat{S}_{yy}(\Omega_\mu,n)$) in order to obtain a second set of correlation function values ($\hat{r}_{yy,g}$ (m,n)), and
determining the speech fundamental frequency estimate ($f_p(n)$) on the basis of the first and second sets of correlation function values ($\hat{r}_{yy_d,g}$ (m,n), $\hat{r}_{yy,g}$ (m, n)).

24. Method (1140) according to claim 23, **characterized in that** the step of determining (1156) the speech fundamental frequency estimate ($f_p(n)$) comprises:

- performing a first frequency-time-transform of the first power density spectrum ($\hat{S}_{yy_d}$ (Ω_μ, n)) in order to obtain a first set of correlation function values ($\hat{r}_{yy_d,g}$ (m, n));
- performing a second frequency-time-transform of the second power density spectrum ($\hat{S}_{yy}(\Omega_\mu,n)$) in order to obtain a second set of correlation function values ($\hat{r}_{\underline{yy,g}}(m,n)$), wherein the step of determining (1156) further comprises determining a set of normalization values ($\bar{\mathbf{S}}_{\tilde{y}\tilde{y}}(\Omega_\mu,n)$) and a set of weighting values ($V(e^{j\Omega_\mu}, n)$) from the second power density spectrum ($S_{yy}(\Omega_\mu, n)$ and using the set of normalization values ($\bar{S}_{yy}$ (Ω_μ,n)) and the set of weighting values *($V(e^{j\Omega_\mu},n)$)* in the first and second frequency-time-transform and wherein the determination of the speech fundamental frequency estimate ($f_p(n)$) is performed on the basis of the first and second sets of correlation function values ($\hat{r}_{yy_d,g}$ (m,n) ,$\hat{r}_{yy,g}(m,n)$).

25. Method (1140) according to claim 24, **characterized in that** the step of determining (1156) the speech fundamental frequency estimate ($f_p(n)$) comprises adaptively compensating the values of the first set of correlation function values ($\hat{r}_{yy_d,g}$ (m, n)) by a correction factor ($\Delta(m,n)$) being based on a value of the second set of correlation function values ($\hat{r}_{yy,g}(m,n)$) in order to obtain a compensated first set of values and determining the speech fundamental frequency estimate ($f_p(n)$) on the basis of the compensated first set of correlation function values ($\hat{r}_{yy_d,g,mod}$ (m,n)) and the second set of correlation function values ($\hat{r}_{\tilde{y}\tilde{y},g}$ (m, n)).

26. Method (1140) according to claim 25, **characterized in that** the step of compensating comprises multiplying the second set of correlation function values ($\hat{r}_{yy,g}$ (m, n)) by a lower bounded quotient between a value of the first set of correlation function values ($\hat{r}_{yy,dg}(m,n)$) and a value of the second set of correlation function values ($\hat{r}_{yy,g}(m, n)$) in order to obtain said compensated first set of correlation function values ($\hat{r}_{\tilde{y}\tilde{y},\, g,\mathrm{mod}}$ (m,n)).

27. Method (1140) according to claim 26, **characterized in that** the step of determining (1156) the speech fundamental frequency estimate ($f_p(n)$) comprises combining the compensated first set of correlation function values ($\hat{r}_{yy,d,\, g,mod}$ (m, n)) and the second set of correlation function values ($\hat{r}_{yy,g}(m,n)$) in order to obtain an extended set of correlation function values ($\hat{r}_{yy,erw}(k,n)$),
wherein the values of the extended set of correlation function values ($\hat{r}_{yy,erw,}$ (k, n)) assume corresponding values from the compensated first set of correlation function values ($\hat{r}_{yy_d,g,mod}$ (m,n)), the second set of correlation function values ($\hat{r}_{yy,g}(m,n)$) or values between the compensated first set of correlation function values ($\hat{r}_{yy_d,g,mod}$ (m,n)) and the second set of correlation function values ($\hat{r}_{yy,g}(m,n)$) and wherein step of determining (1156) the speech fundamental frequency estimate ($f_p(n)$) further comprises determining the speech fundamental frequency estimate ($f_p(n)$) on the basis of said extended set of correlation function values ($\hat{r}_{yy,erw}(k,n)$).

28. Method (1140) according to one of claims 23 to 27, **characterized in that** the step of determining (1156) the speech fundamental frequency estimate ($f_p(n)$) comprises determining the speech fundamental frequency estimate ($f_p(n)$) by searching the index of a maximum value ($\tau_p(n)$) from the extended set of correlation function values ($\hat{r}_{yy,erw}(k,n)$) within a predetermined number of indices (k) of the values of the extended set of correlation values ($\hat{r}_{yy,erw}(k,n)$), from the first or second set of correlation function values ($\hat{r}_{yyd,g}(m, n)$, $\hat{r}_{yy,g}(m, n)$) within a predetermined number of indexes (m) of values of the first respectively second set of correlation function values ($\hat{r}_{yyd,g}(m, n)$, $\hat{r}_{yy,g}(m, n)$) or from the compensated first set of correlation function values ($\hat{r}_{yy,d,g,mod}(m, n)$) within the predetermined number of indices (m) of values of the compensated first set of correlation function values ($\hat{r}_{yy\ d,g,mod}(m,n)$) and wherein the step of determining (1156) the speech fundamental frequency estimate ($f_p(n)$) furthermore comprises determining the speech fundamental frequency estimate ($f_p(n)$) as the product of a sampling frequency (f_s) and a reciprocal value of said searched index ($\tau_p(n)$).

29. Method (1140) according to claim 28, **characterized in that** the step of determining (1156) the speech fundamental frequency estimate ($f_p(n)$) comprises determining a reliability factor ($p_{f_p}(n)$) for the determined speech fundamental frequency estimate ($f_p(n)$) and for blocking an output of the determined speech fundamental frequency estimate ($f_p(n)$) in the case the determined reliability factor ($p_{f_p}(n)$) for the determined speech fundamental frequency estimate ($f_p(n)$) is below predetermined reliability factor (p_0).

30. Method (1140) according to claim 29, **characterized in that** the step of determining (1156) the speech fundamental frequency estimate ($f_p(n)$) comprises determining said reliability factor ($p_{f_p}(n)$) by dividing the maximum value ($\tilde{\tau}_p(n)$) at said searched by the first value of the extended set of correlation function values ($\hat{r}_{yy,erw}(k,n)$) or, respectively the first, the compensated first or second set of correlation function values ($\hat{r}_{yyd,g}(m,n)$, $\hat{r}_{yyd,g,mod}(m, n)$, $\hat{r}_{yy,g}(m, n)$).

31. Method (1140) according to one of claims 23 to 30 and claim 24, **characterized in that** the step of providing values of a second power density spectrum ($\hat{S}_{\tilde{y}\tilde{y}}(\Omega_\mu, n)$) comprises determining an estimate of the power density spectrum of background noise ($\hat{S}_{nn}(\Omega_\mu, n)$) and determining a noise suppression factor ($V(e^{j\Omega_\mu},n)$) on the basis of said power density spectrum of background noise ($\hat{S}_{nn}(\Omega_\mu, n)$), and the step of determining (1156) the speech fundamental frequency estimate ($f_p(n)$) comprises multiplying the first and second power density spectrum with said noise suppression factor ($V(e^{j\Omega_\mu}, n)$) prior to the frequency-time-transform of the first respectively second power density spectrum ($\hat{S}_{yyd}(\Omega_\mu,n)$, $\hat{S}_{yy}(\Omega_\mu,n)$).

32. Method (1140) according to claim 31, **characterized in that** the step of providing values of a second power density spectrum ($\hat{S}_{yy}(\Omega_\mu,n)$) comprises determining the noise suppression factor as the maximum of a predetermined maximum suppression coefficient (V_0) and a term being dependent on a quotient of the estimate of the power density spectrum of background noise ($\hat{S}_{nn}(\Omega_\mu, n)$) and the second power density spectrum ($\hat{S}_{yy}(\Omega_\mu, n)$).

33. Method (1140) according to claim 32, **characterized in that** the step of providing values of a second power density spectrum ($\hat{S}_{\tilde{y}\tilde{y}}(\Omega_\mu, n)$) comprises determining the estimate of the power density spectrum of background noise ($\hat{S}_{nn}(\Omega_\mu, n)$) in speech pauses or for determining the estimate of the power density spectrum of background noise ($\hat{S}_{nn}(\Omega_\mu,n)$) from a segment-wise estimation of the minima of the power of a differential signal.

34. Method (1140) according to one of claims 31 to 33, **characterized in that** the noise suppression factor is defined by

$$V\left(e^{j\Omega_\mu}, n\right) = \max\left\{V_0,\ 1 - \beta\frac{\hat{S}_{nn}(\Omega_\mu, n)}{\hat{S}_{yy}(\Omega_\mu, n)}\right\}$$

wherein $\hat{S}_{nn}(\Omega_\mu, n)$ denotes the estimate of the power density spectrum of the background noise, $\hat{S}_{yy}(\Omega_\mu, n)$ denotes the second power density spectrum, V_0 denotes a predefined maximum attenuation factor and β denotes a value for overestimating the power density spectrum of the background noise ($\hat{S}_{nn}(\Omega_\mu,n)$).

35. Method (1140) according to one of claims 24 to 34, **characterized in that** the step of determining (1156) the speech fundamental frequency estimate ($f_p(n)$) comprises reestimating the speech fundamental frequency estimate ($f_p(n)$) in the case the determined speech fundamental frequency estimate is below the predefined frequency value (f_k) wherein the step of determining (1156) the speech fundamental frequency estimate ($f_p(n)$) comprises performing the reestimation by searching a further index (k, m) of a further maximum value ($\hat{r}_p(n)$) of the extended set of correlation function values ($\hat{r}_{yy,erw}(k,n)$), the first or second set of correlation function values ($\hat{r}_{yy_d,g}(m,n)$,$\hat{r}_{yy,g}(m,n)$) or the compensated first set of correlation function values ($\hat{r}_{yy_d,g,mod}(m,n)$) within a further number of values of said sets of correlation function values and outputing a product of a sampling frequency (f_s) and a reciprocal value of said further index ($\tilde{\tau}_p(n)$) as the determined speech fundamental frequency estimate.

36. Method (1140) according to claim 35, **characterized in that** the step of determining (1156) the speech fundamental frequency estimate ($f_p(n)$) comprises searching said index (k, m) of said further maximum value ($\tilde{\tau}_p(n)$) using a number of values k of said sets of correlation function values which is defined by

$$\frac{f_s}{f_{\mathrm{p,max}}} \leq k < \frac{f_s}{2f_{\mathrm{p}}(n)} + k_0$$

wherein k denotes the number of values of said sets of correlation function values, $f_p(n)$ denotes the previously determined speech fundamental frequency estimate, $f_{p,max}$ denotes a predefined value of a maximal possible speech fundamental frequency, f_s denotes a sampling frequency and ko denotes a constant.

37. Method (1140) according to one of claims 35 or 36, **characterized in that** the step of determining (1156) the speech fundamental frequency estimate ($f_p(n)$) comprises outputing said product as the predetermined speech fundamental frequency estimate ($f_p(n)$) only in the case that the further index ($\tilde{\tau}_p(n)$) is larger than 60 percent of the previously searched maximal index ($\tau_p(n)$) as well as the value ($\hat{r}_{yy,erw}(\tilde{\tau}_p(n), n)$) of the extended set of correlation function values ($\hat{r}_{yy,erw}(k,n)$) at said further index ($\tilde{\tau}_p(n)$) is larger than a previously defined amplitude value ($\tilde{p}_0$).

38. Method (1140) according to one of claims 24 to 37, **characterized in that** the step of determining the speech fundamental frequency estimate ($f_p(n)$) comprises modifying a speech fundamental period ($\tilde{\tau}_p(n)$) corresponding to said determined speech fundamental frequency estimate ($f_p(n)$) by a interpolation correction term ($\Delta_p(n)$) prior of outputing said speech fundamental frequency estimate (fp(n)), wherein said interpolation correction term ($\Delta_p(n)$) is dependent on values of said first or second set of correlation function values ($\hat{r}_{yy_d,g}(m,n)$, $\hat{r}_{yy,g}(m, n)$), of said extended set of correlation function values ($\hat{r}_{yy,erw}(k,n)$) or said compensated first set of correlation function values ($\hat{r}_{yy_d,g,\mathrm{mod}}(m,n)$), respectively.

39. Method (1140) according to one of the preceding claims, **characterized in that** the method further comprises a step of receiving the frequency domain versions (Y_1, Y_2) of the first and second set of time domain signal values (y_1, y_2), frequency domain filtering said frequency domain versions in order to obtain said first and second sets of values ($\tilde{Y}_1$, $\tilde{Y}_2$), respectively, and providing said first and second sets of values ($\tilde{Y}_1$, $\tilde{Y}_2$) the first and second power density spectrum calculator respectively.

40. Method (1140) according to claim 39, **characterized in that** the step of frequency domain filtering is only performed for frequencies below a predefined limiting frequency.

41. Method (1140) according to claim 40, **characterized in that** the step of frequency domain filtering comprises delaying values of said frequency domain versions being above said predefined limiting frequency.

42. Computer program product having a program code for performing the method according to one of claims 23 to 41, when the computer program runs on a computer.

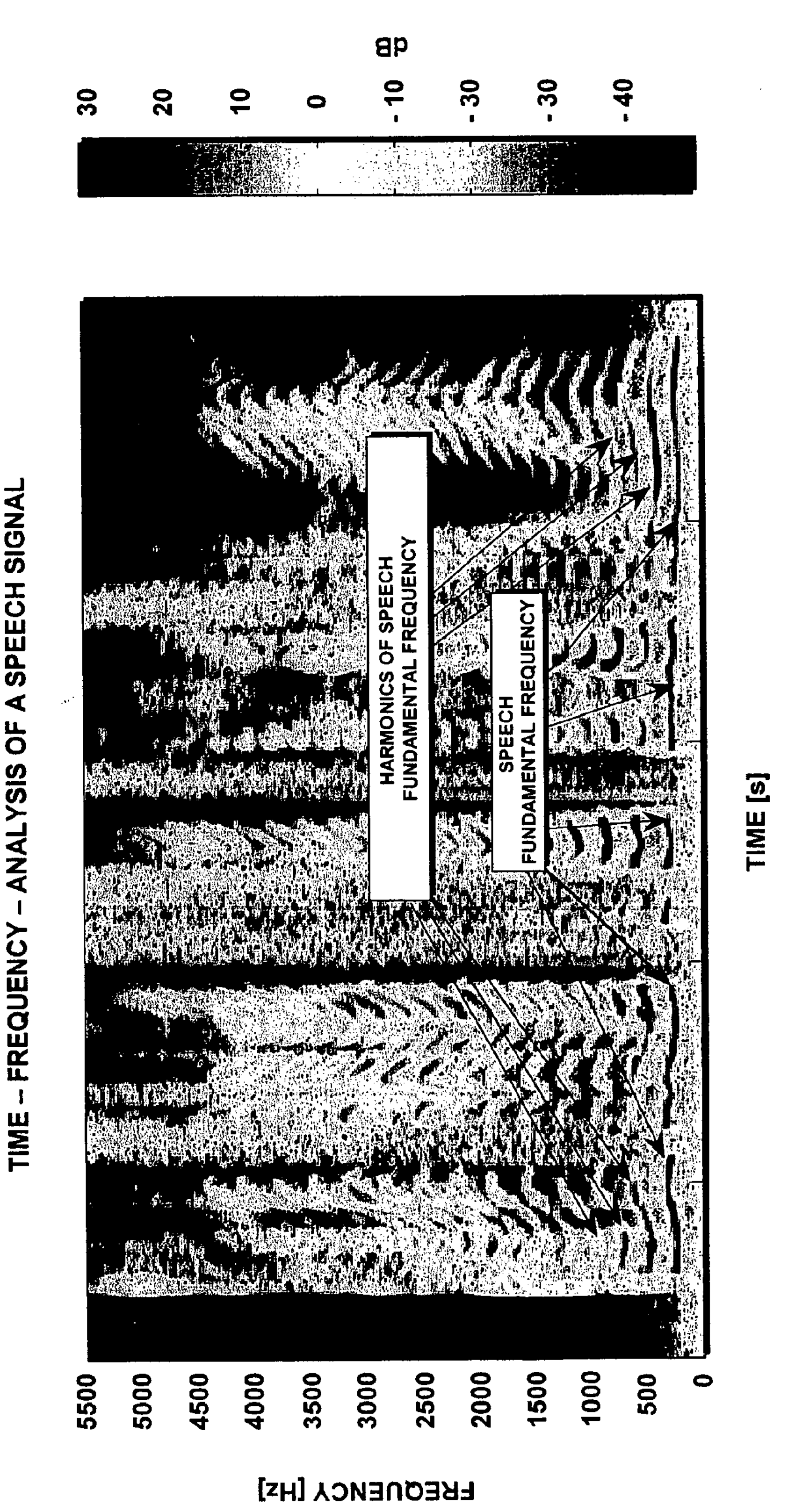

TIME – FREQUENCY – ANALYSIS OF A SPEECH SIGNAL
FREQUENCY [Hz]
5500
5000
4500
4000
3500
3000
2500
2000
1500
1000
500
0
TIME [s]
HARMONICS OF SPEECH FUNDAMENTAL FREQUENCY
SPEECH FUNDAMENTAL FREQUENCY
30
20
10
0
- 10
- 30
- 30
- 40
dB

FIG. 1

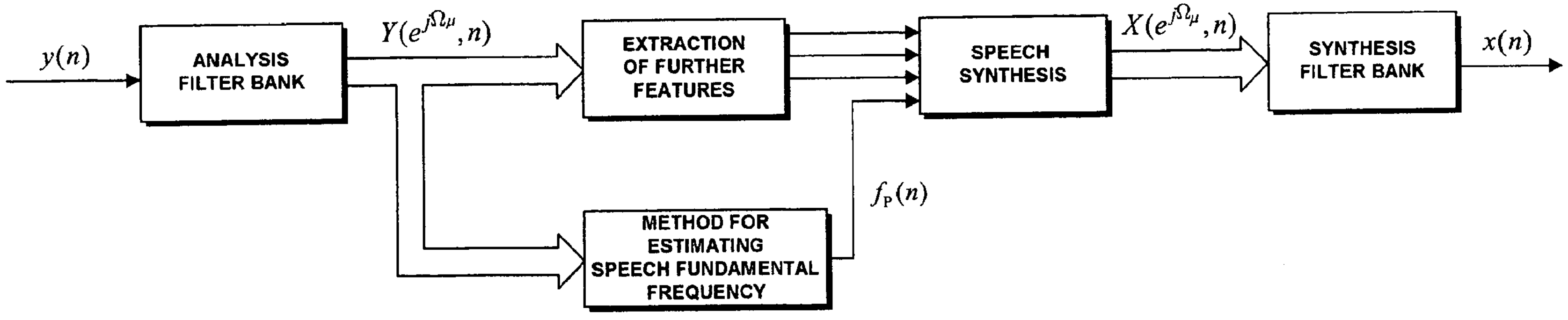

$y(n)$
ANALYSIS FILTER BANK
$Y(e^{j\Omega_\mu},n)$
EXTRACTION OF FURTHER FEATURES
METHOD FOR ESTIMATING SPEECH FUNDAMENTAL FREQUENCY
$f_P(n)$
SPEECH SYNTHESIS
$X(e^{j\Omega_\mu},n)$
SYNTHESIS FILTER BANK
$x(n)$

FIG. 2

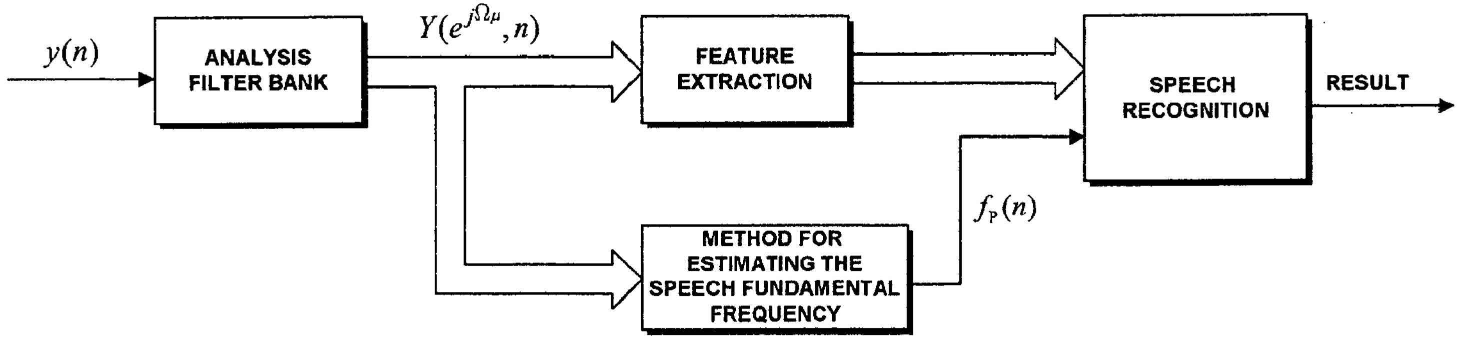

$y(n)$
ANALYSIS FILTER BANK
$Y(e^{j\Omega_\mu},n)$
FEATURE EXTRACTION
METHOD FOR ESTIMATING THE SPEECH FUNDAMENTAL FREQUENCY
$f_P(n)$
SPEECH RECOGNITION
RESULT

FIG. 3

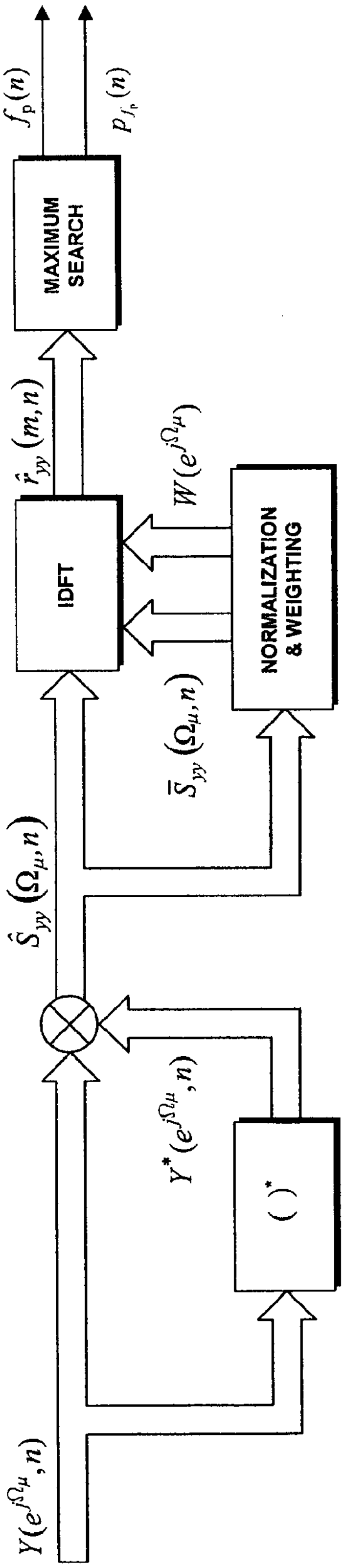
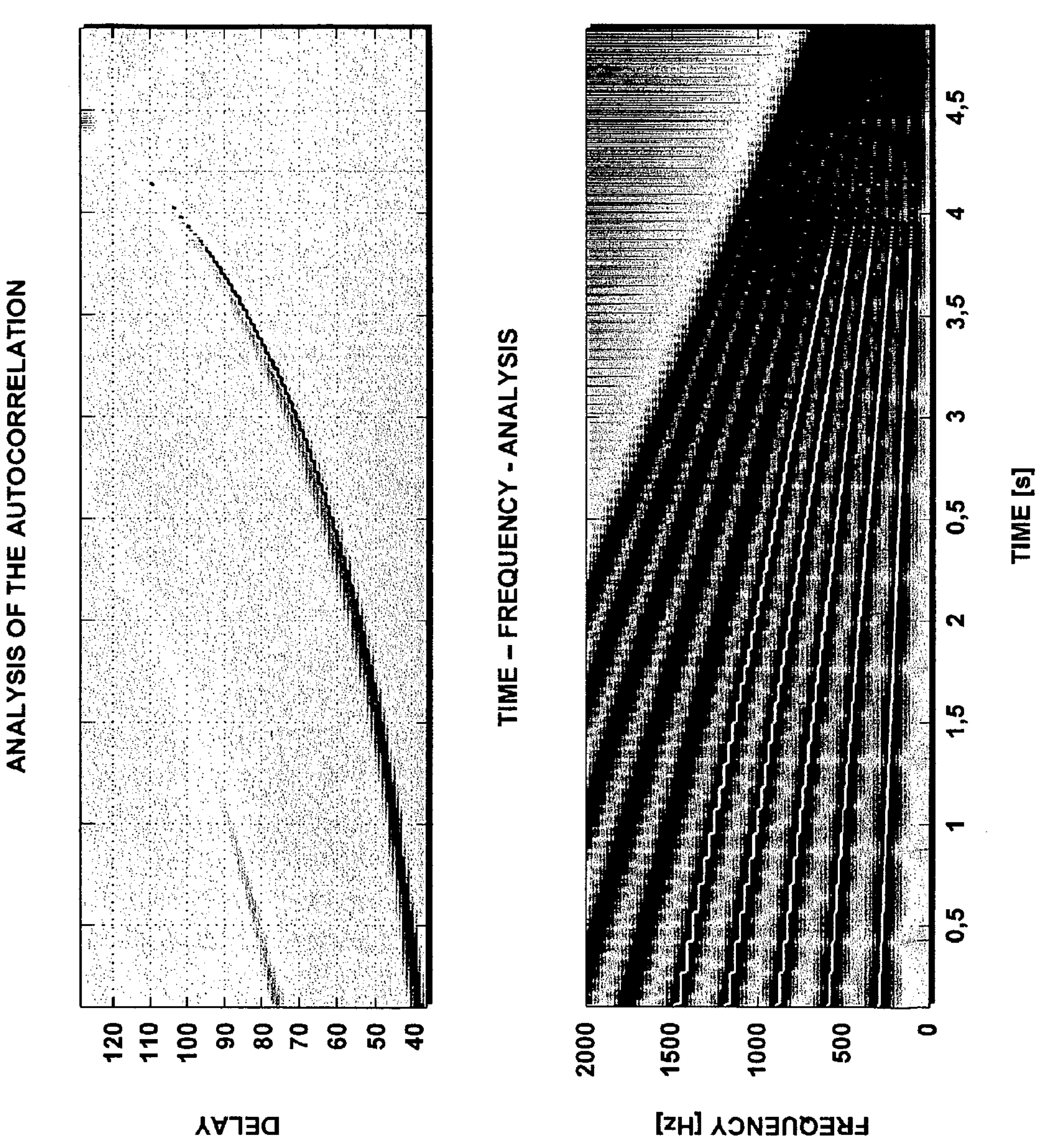

$Y(e^{j\Omega_\mu},n)$
$(\ )^*$
$Y^*(e^{j\Omega_\mu},n)$
$\hat{S}_{yy}(\Omega_\mu,n)$
$\bar{S}_{yy}(\Omega_\mu,n)$
NORMALIZATION & WEIGHTING
$W(e^{j\Omega_\mu})$
IDFT
$\hat{r}_{yy}(m,n)$
MAXIMUM SEARCH
$f_p(n)$
$p_{f_p}(n)$

FIG. 4



ANALYSIS OF THE AUTOCORRELATION

TIME – FREQUENCY - ANALYSIS

FIG. 5

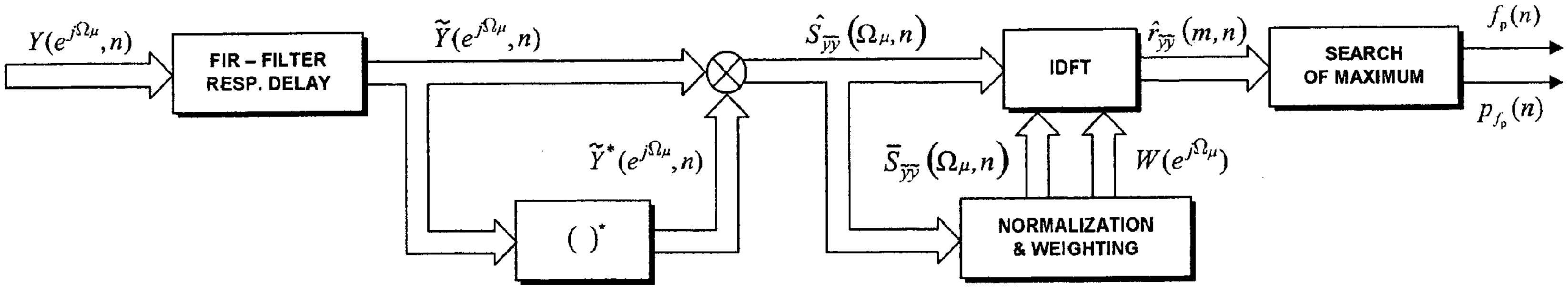

$Y(e^{j\Omega_\mu},n)$
FIR – FILTER RESP. DELAY
$\tilde{Y}(e^{j\Omega_\mu},n)$
$\tilde{Y}^*(e^{j\Omega_\mu},n)$
$(\ )^*$
$\hat{S}_{\tilde{y}\tilde{y}}(\Omega_\mu,n)$
$\bar{S}_{\tilde{y}\tilde{y}}(\Omega_\mu,n)$
NORMALIZATION & WEIGHTING
$W(e^{j\Omega_\mu})$
IDFT
$\hat{r}_{\tilde{y}\tilde{y}}(m,n)$
SEARCH OF MAXIMUM
$f_p(n)$
$p_{f_p}(n)$

FIG. 6

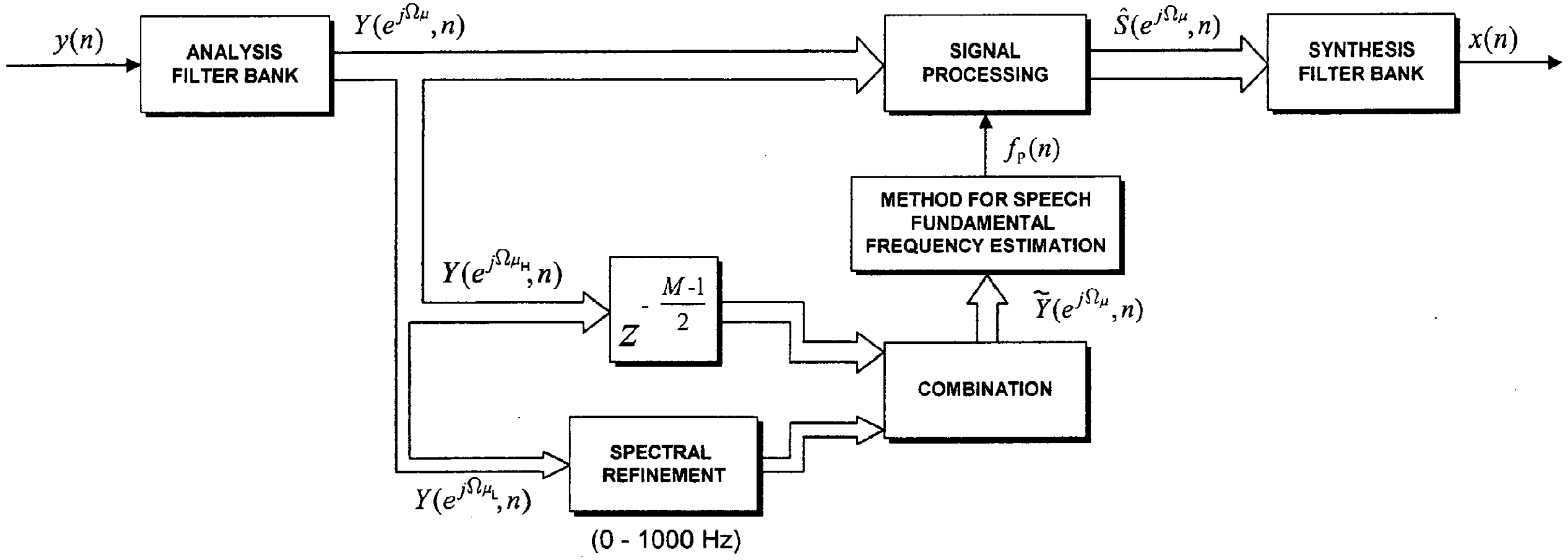
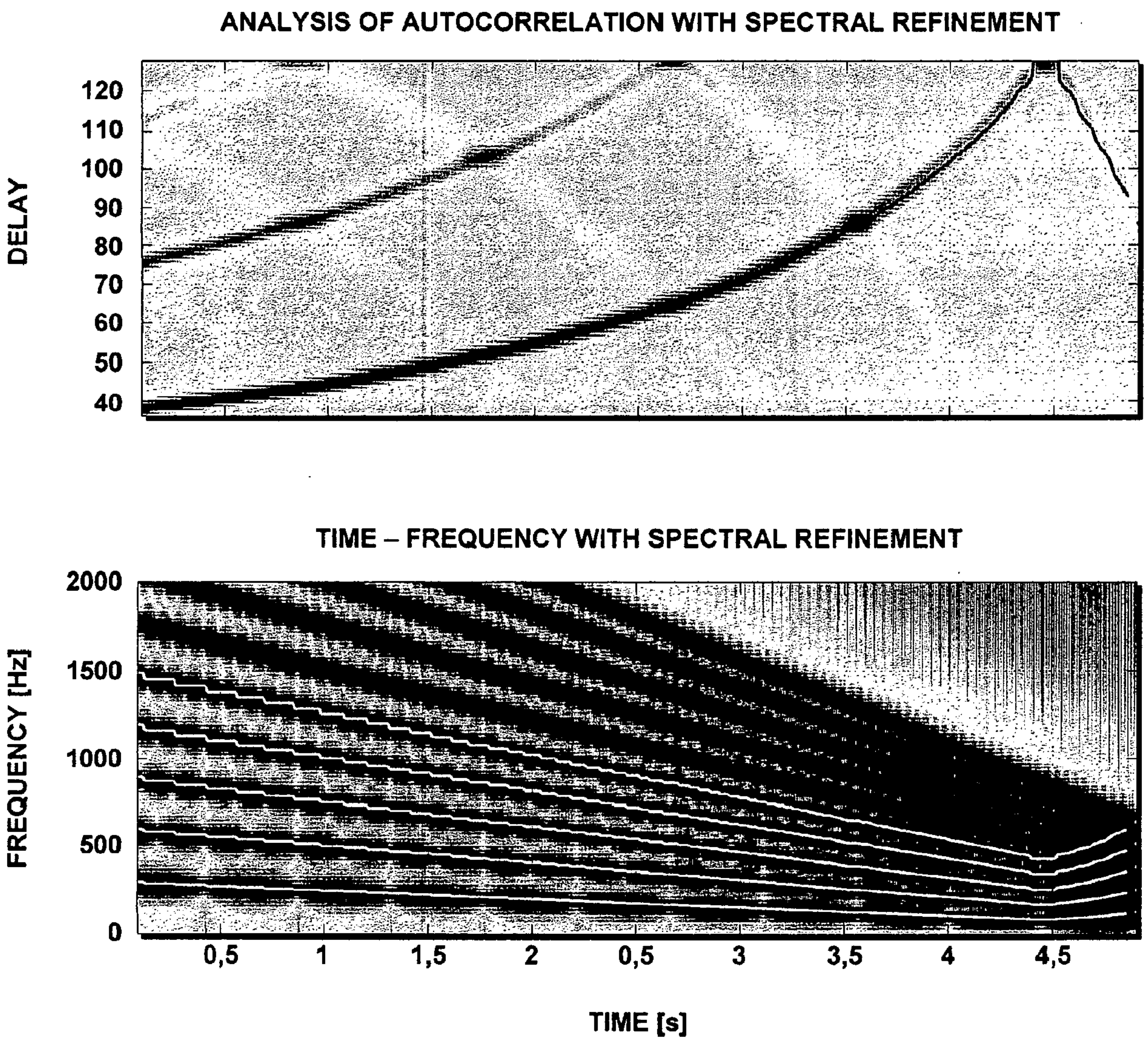

$y(n)$
ANALYSIS FILTER BANK
$Y(e^{j\Omega_\mu},n)$
SIGNAL PROCESSING
$\hat{S}(e^{j\Omega_\mu},n)$
SYNTHESIS FILTER BANK
$x(n)$
$f_P(n)$
METHOD FOR SPEECH FUNDAMENTAL FREQUENCY ESTIMATION
$Y(e^{j\Omega_{\mu_H}},n)$
$Z^{-\frac{M-1}{2}}$
$\tilde{Y}(e^{j\Omega_\mu},n)$
COMBINATION
SPECTRAL REFINEMENT
$Y(e^{j\Omega_{\mu_L}},n)$
(0 - 1000 Hz)

FIG. 7

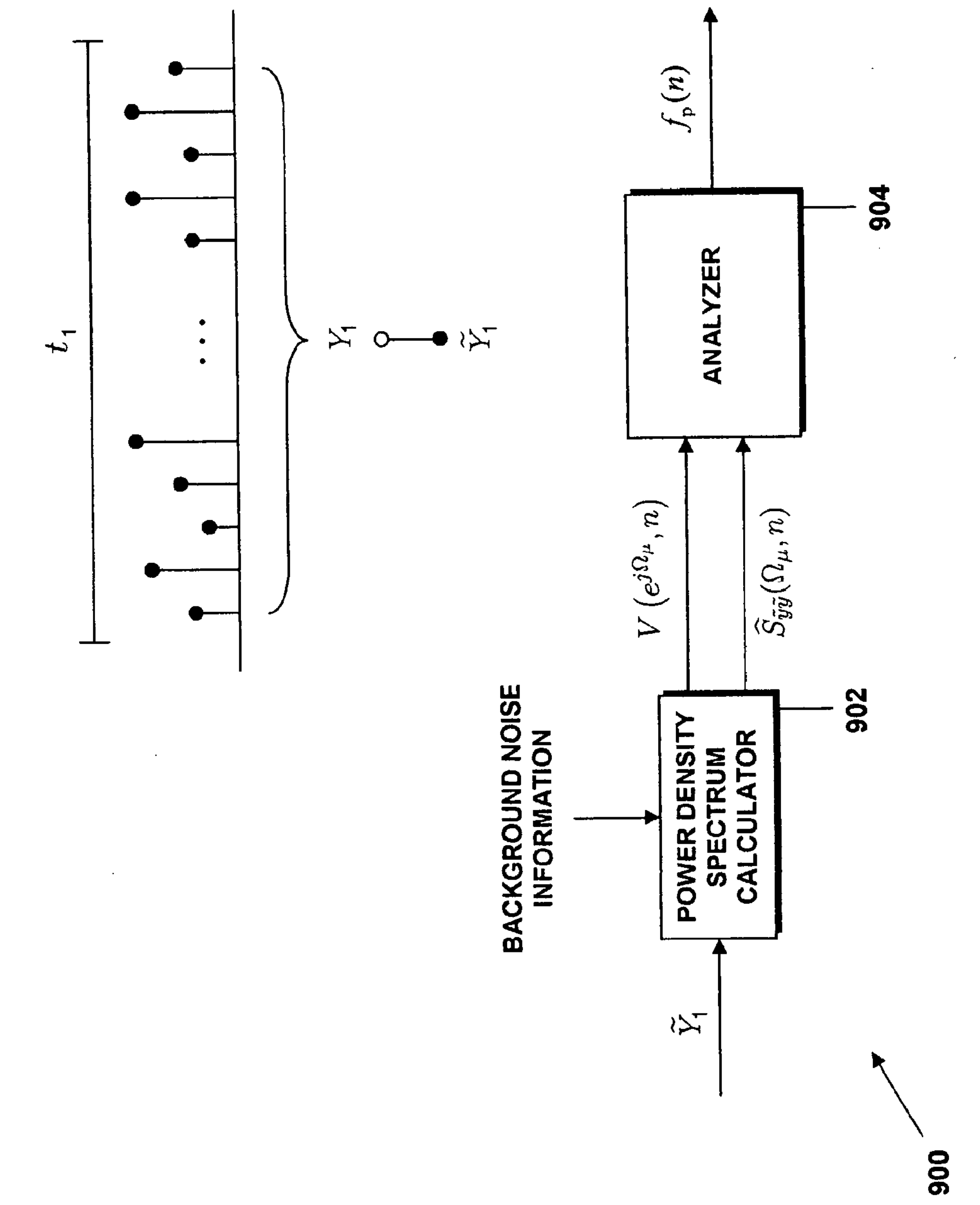


ANALYSIS OF AUTOCORRELATION WITH SPECTRAL REFINEMENT

TIME – FREQUENCY WITH SPECTRAL REFINEMENT

FIG. 8

FIG. 9A


t_1
Y_1
$\tilde{Y}_1$
BACKGROUND NOISE INFORMATION
$\tilde{Y}_1$
POWER DENSITY SPECTRUM CALCULATOR
902
$V\left(e^{j\Omega_\mu}, n\right)$
$\widehat{S}_{\tilde{y}\tilde{y}}(\Omega_\mu, n)$
ANALYZER
904
$f_p(n)$
900

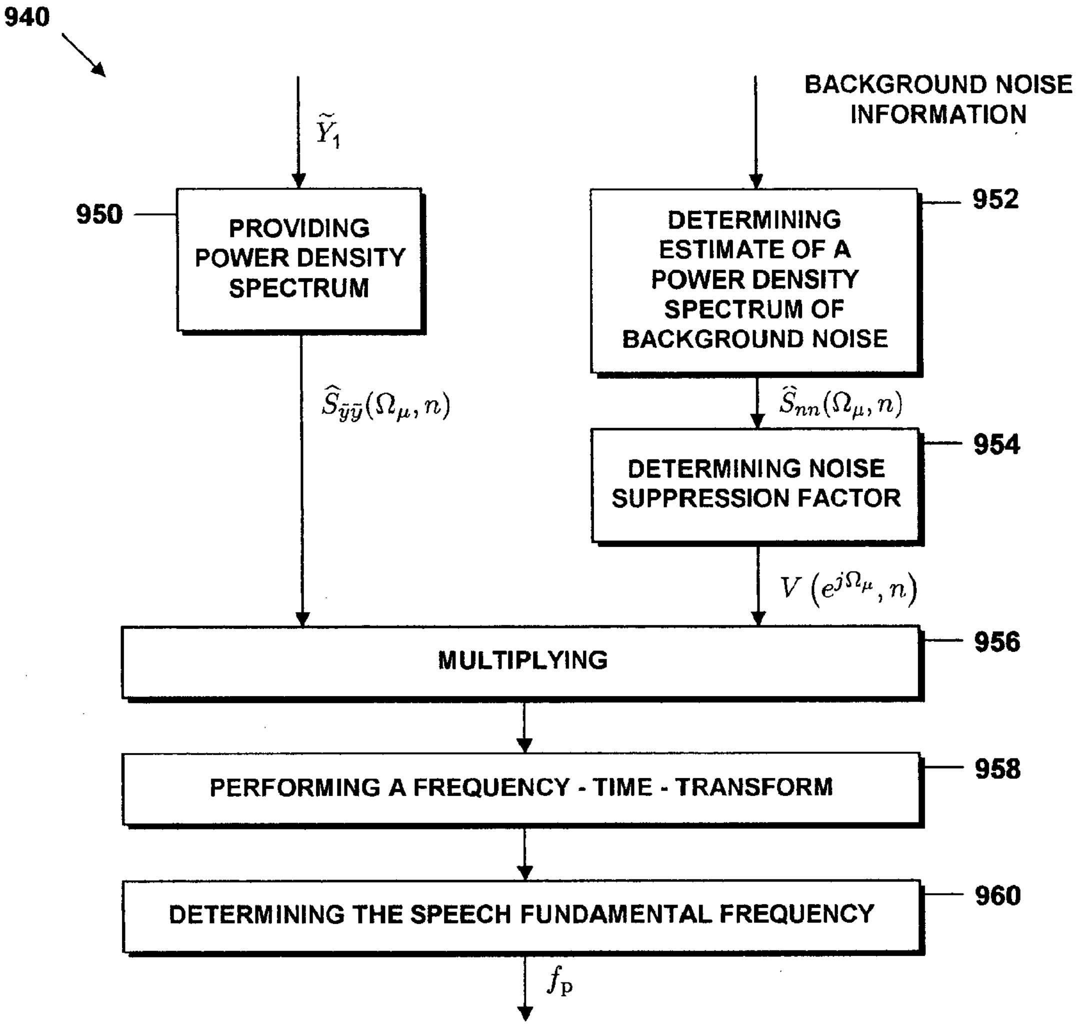
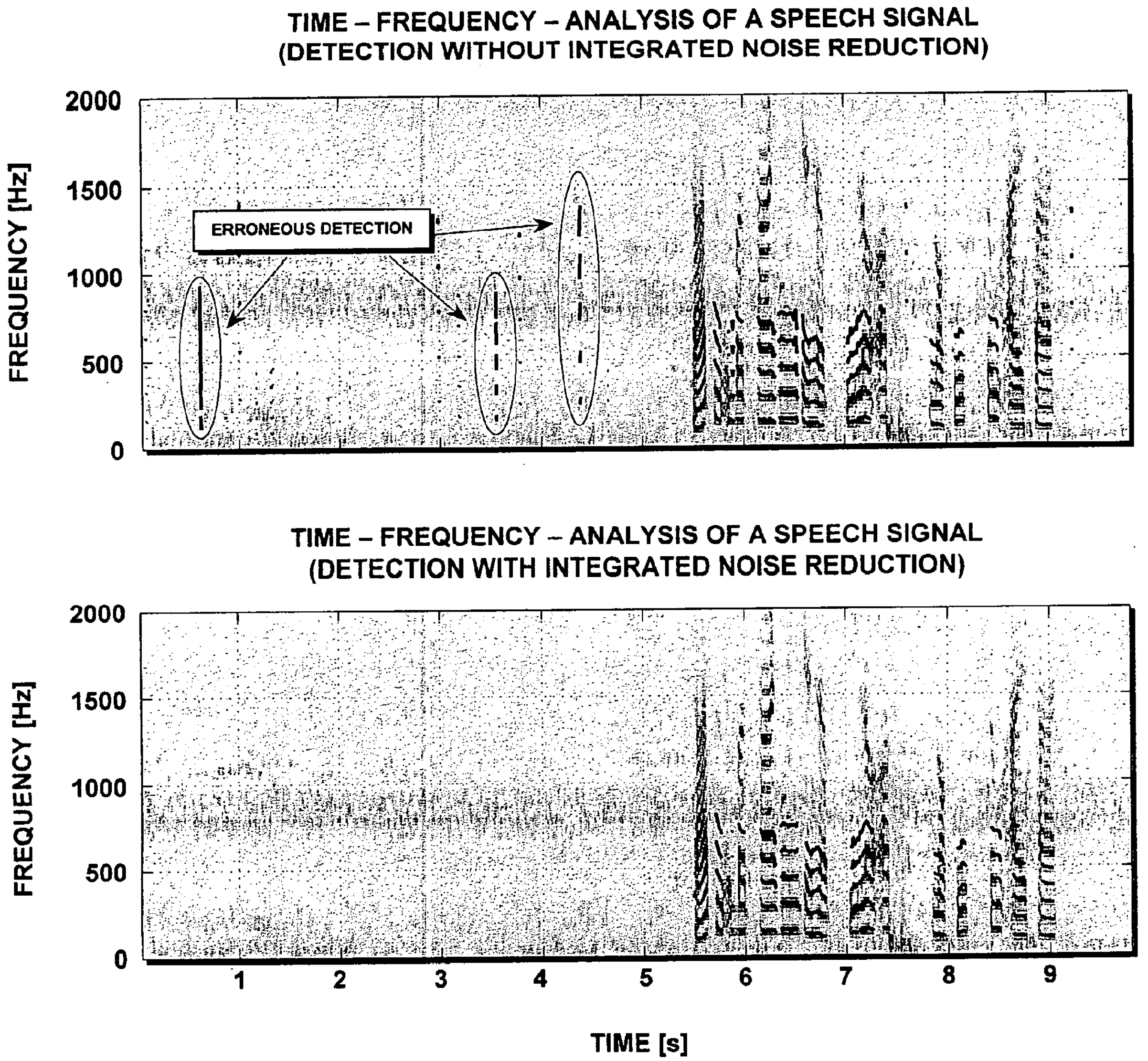

940
$\widetilde{Y}_1$
BACKGROUND NOISE INFORMATION
950
PROVIDING POWER DENSITY SPECTRUM
952
DETERMINING ESTIMATE OF A POWER DENSITY SPECTRUM OF BACKGROUND NOISE
$\widehat{S}_{\tilde{y}\tilde{y}}(\Omega_\mu, n)$
$\widehat{S}_{nn}(\Omega_\mu, n)$
954
DETERMINING NOISE SUPPRESSION FACTOR
$V\left(e^{j\Omega_\mu}, n\right)$
956
MULTIPLYING
958
PERFORMING A FREQUENCY - TIME - TRANSFORM
960
DETERMINING THE SPEECH FUNDAMENTAL FREQUENCY
f_p

FIG. 9B



TIME – FREQUENCY – ANALYSIS OF A SPEECH SIGNAL
(DETECTION WITHOUT INTEGRATED NOISE REDUCTION)

TIME – FREQUENCY – ANALYSIS OF A SPEECH SIGNAL
(DETECTION WITH INTEGRATED NOISE REDUCTION)

FIG. 10

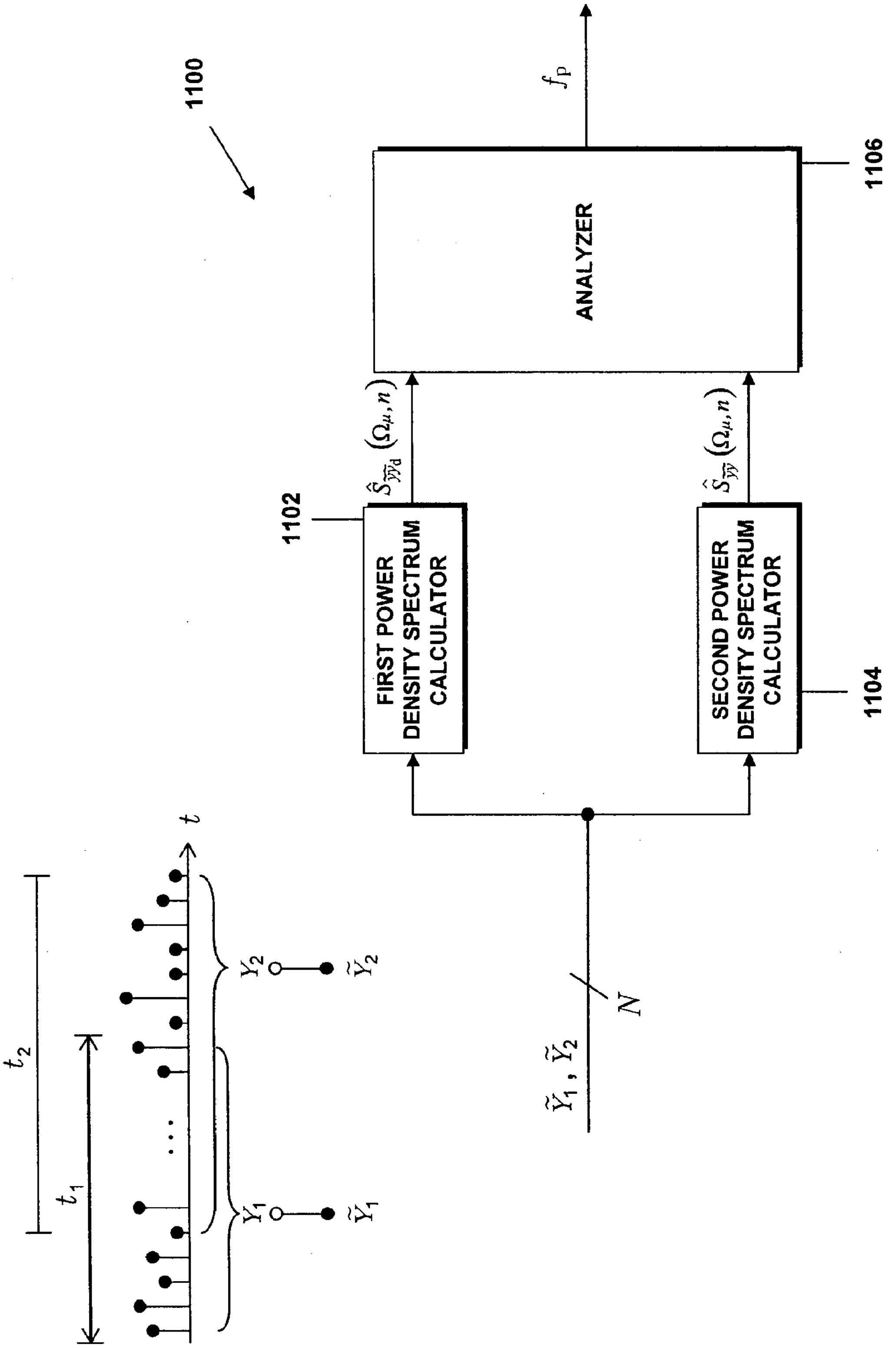

1100
t_2
t_1
t
...
Y_1
Y_2
$\tilde{Y}_1$
$\tilde{Y}_2$
$\tilde{Y}_1, \tilde{Y}_2$
N
1102
FIRST POWER DENSITY SPECTRUM CALCULATOR
1104
SECOND POWER DENSITY SPECTRUM CALCULATOR
$\hat{S}_{\tilde{y}\tilde{y}_d}(\Omega_\mu, n)$
$\hat{S}_{\tilde{y}\tilde{y}}(\Omega_\mu, n)$
ANALYZER
1106
f_p

FIG. 11A

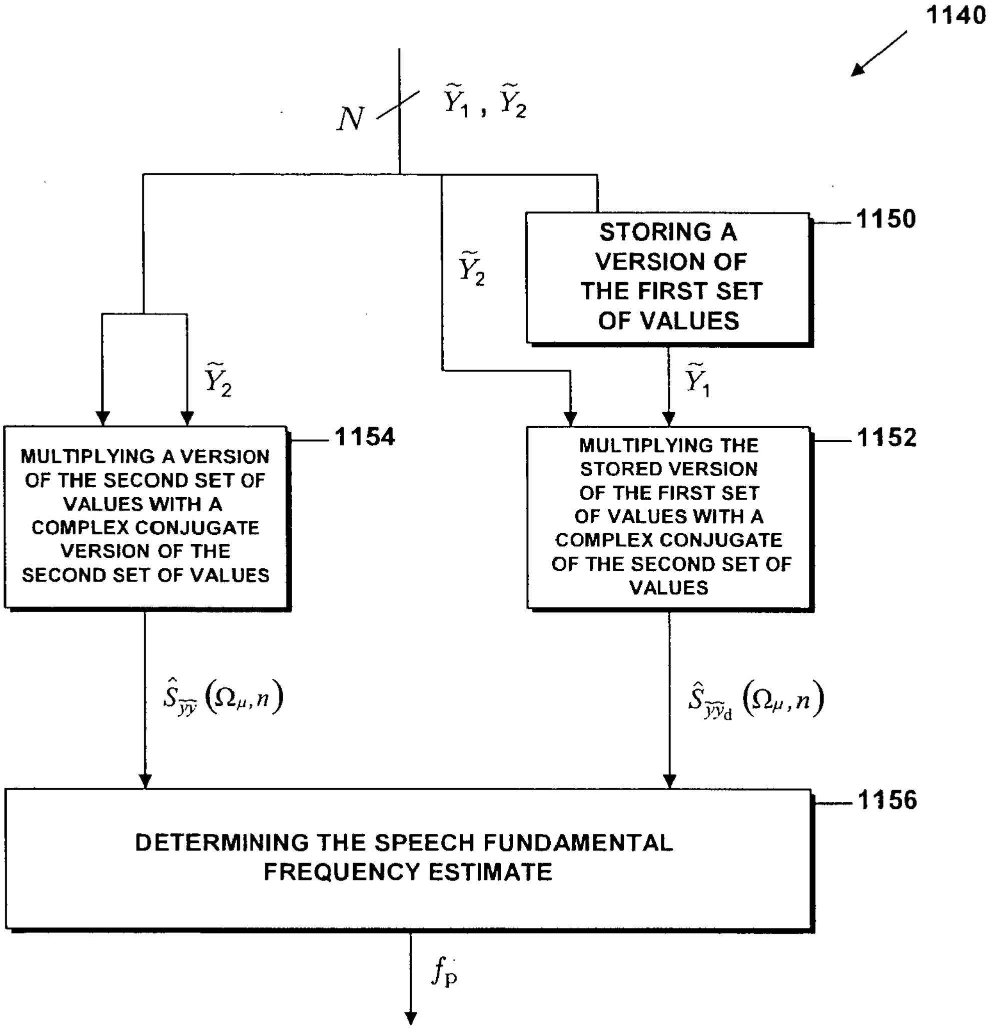

1140
N
$\tilde{Y}_1$, $\tilde{Y}_2$
$\tilde{Y}_2$
STORING A VERSION OF THE FIRST SET OF VALUES
1150
$\tilde{Y}_2$
$\tilde{Y}_1$
MULTIPLYING A VERSION OF THE SECOND SET OF VALUES WITH A COMPLEX CONJUGATE VERSION OF THE SECOND SET OF VALUES
1154
MULTIPLYING THE STORED VERSION OF THE FIRST SET OF VALUES WITH A COMPLEX CONJUGATE OF THE SECOND SET OF VALUES
1152
$\hat{S}_{\tilde{y}\tilde{y}}(\Omega_\mu,n)$
$\hat{S}_{\tilde{y}\tilde{y}_d}(\Omega_\mu,n)$
DETERMINING THE SPEECH FUNDAMENTAL FREQUENCY ESTIMATE
1156
f_p

FIG. 11B

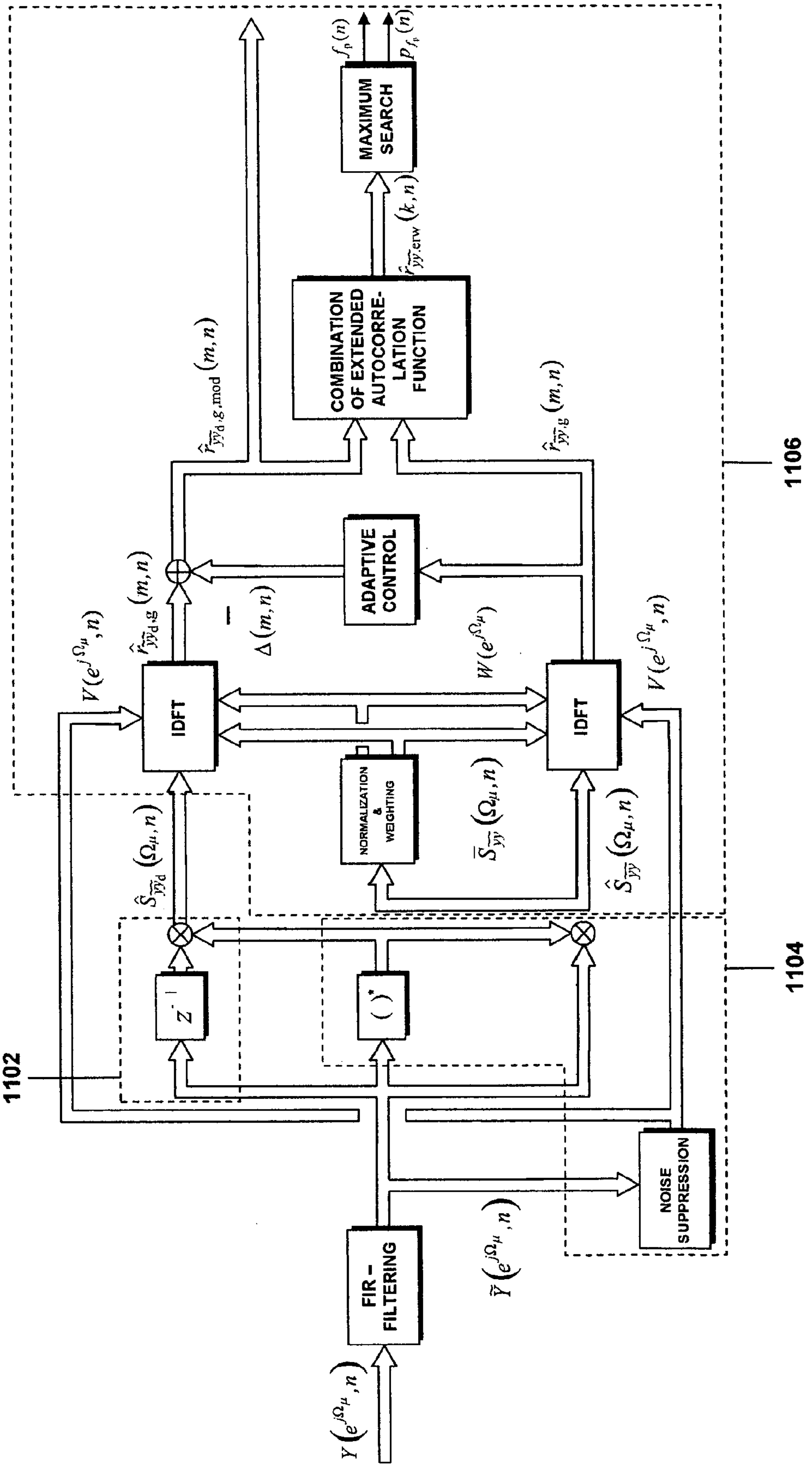

$Y\left(e^{j\Omega_\mu},n\right)$
FIR - FILTERING
$\tilde{Y}\left(e^{j\Omega_\mu},n\right)$
NOISE SUPPRESSION
1102
z^{-1}
$(\ )^*$
1104
$\hat{S}_{\tilde{y}\tilde{y}_d}(\Omega_\mu,n)$
NORMALIZATION & WEIGHTING
$\bar{S}_{\tilde{y}\tilde{y}}(\Omega_\mu,n)$
$\hat{S}_{\tilde{y}\tilde{y}}(\Omega_\mu,n)$
$V(e^{j\Omega_\mu},n)$
IDFT
IDFT
$W(e^{j\Omega_\mu})$
$V(e^{j\Omega_\mu},n)$
$\hat{r}_{\tilde{y}\tilde{y}_d,g}(m,n)$
$-$
$\Delta(m,n)$
ADAPTIVE CONTROL
1106
$\hat{r}_{\tilde{y}\tilde{y}_d,g,\mathrm{mod}}(m,n)$
$\hat{r}_{\tilde{y}\tilde{y},g}(m,n)$
COMBINATION OF EXTENDED AUTOCORRE-LATION FUNCTION
$\hat{r}_{\tilde{y}\tilde{y},\mathrm{erw}}(k,n)$
MAXIMUM SEARCH
$f_p(n)$
$p_{f_p}(n)$

FIG. 12

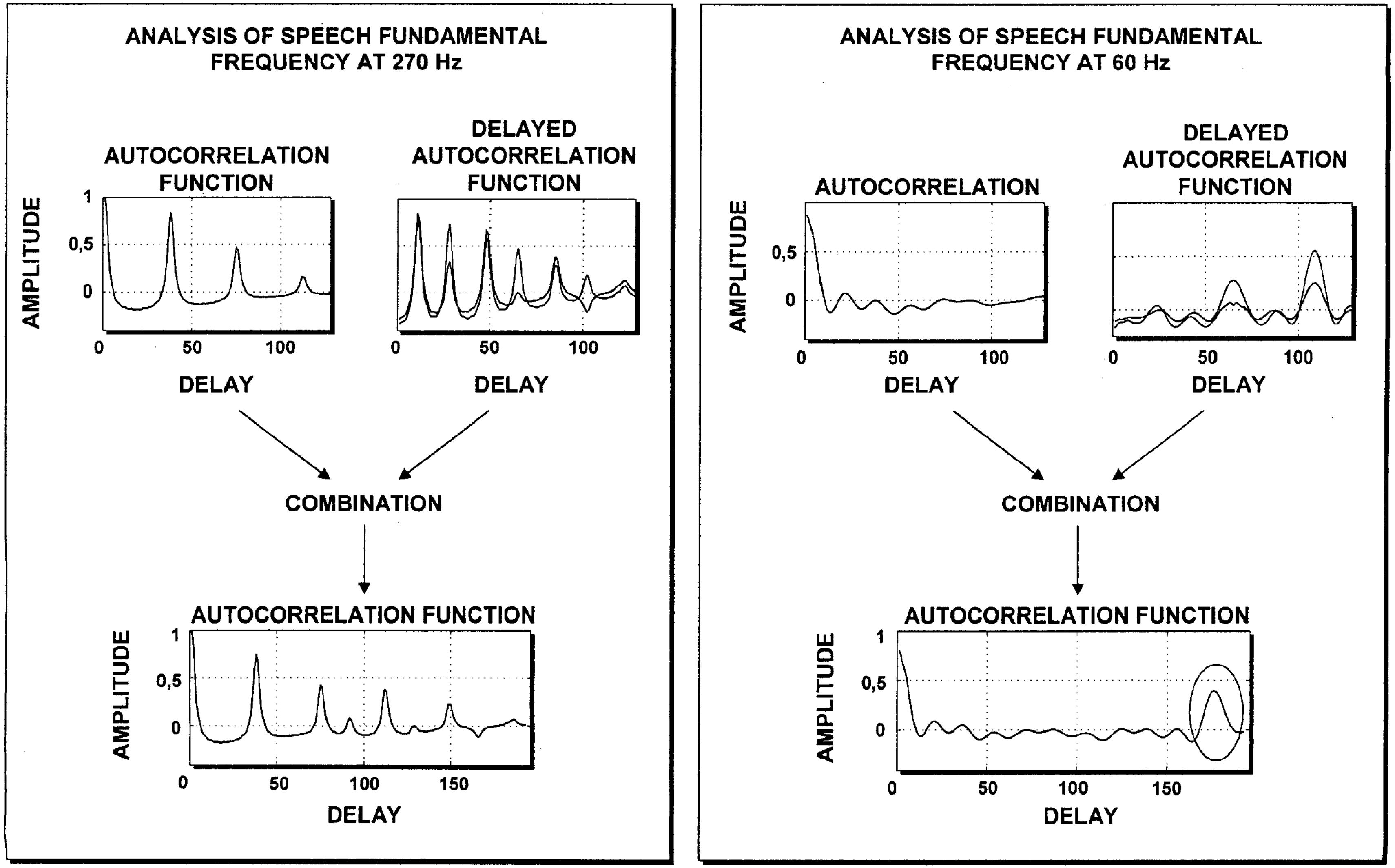
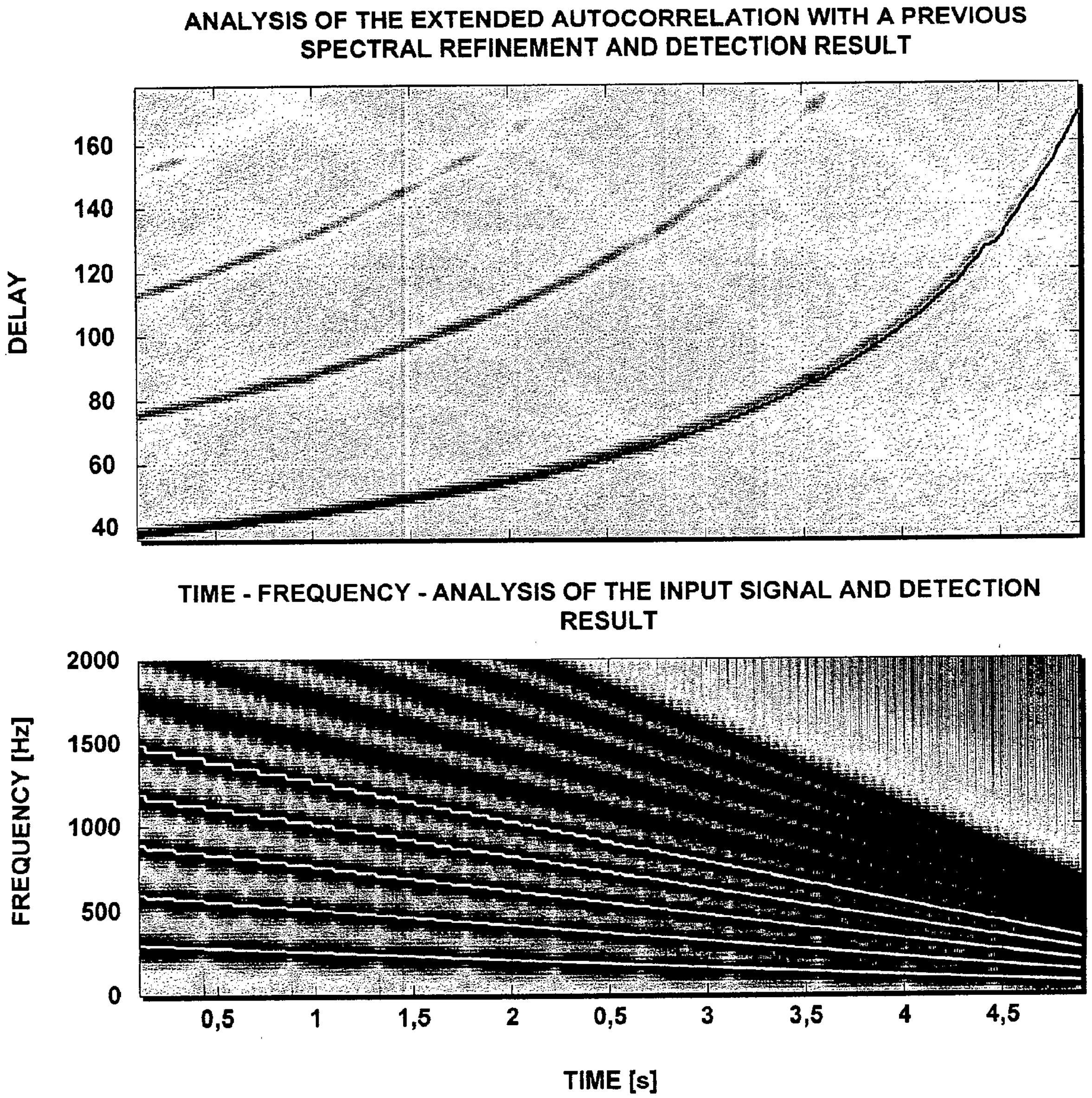

ANALYSIS OF SPEECH FUNDAMENTAL FREQUENCY AT 270 Hz
AUTOCORRELATION FUNCTION
DELAYED AUTOCORRELATION FUNCTION
AMPLITUDE
1
0,5
0
0
50
100
DELAY
DELAY
COMBINATION
AUTOCORRELATION FUNCTION
AMPLITUDE
1
0,5
0
0
50
100
150
DELAY
ANALYSIS OF SPEECH FUNDAMENTAL FREQUENCY AT 60 Hz
AUTOCORRELATION
DELAYED AUTOCORRELATION FUNCTION
AMPLITUDE
0,5
0
0
50
100
DELAY
DELAY
COMBINATION
AUTOCORRELATION FUNCTION
AMPLITUDE
1
0,5
0
0
50
100
150
DELAY

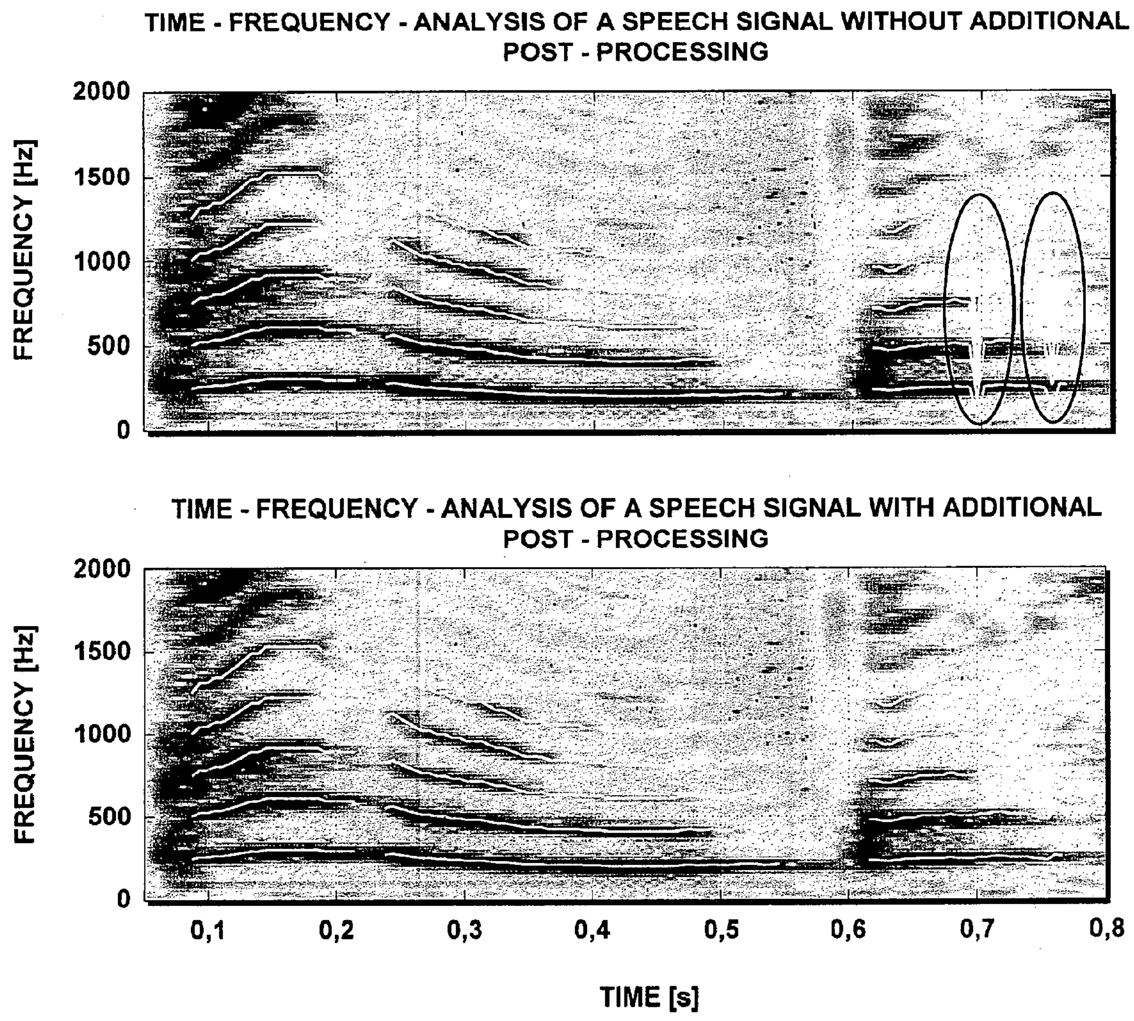
FIG. 13



ANALYSIS OF THE EXTENDED AUTOCORRELATION WITH A PREVIOUS SPECTRAL REFINEMENT AND DETECTION RESULT

TIME - FREQUENCY - ANALYSIS OF THE INPUT SIGNAL AND DETECTION RESULT

FIG. 14



TIME - FREQUENCY - ANALYSIS OF A SPEECH SIGNAL WITHOUT ADDITIONAL POST - PROCESSING

TIME - FREQUENCY - ANALYSIS OF A SPEECH SIGNAL WITH ADDITIONAL POST - PROCESSING

FIG. 15

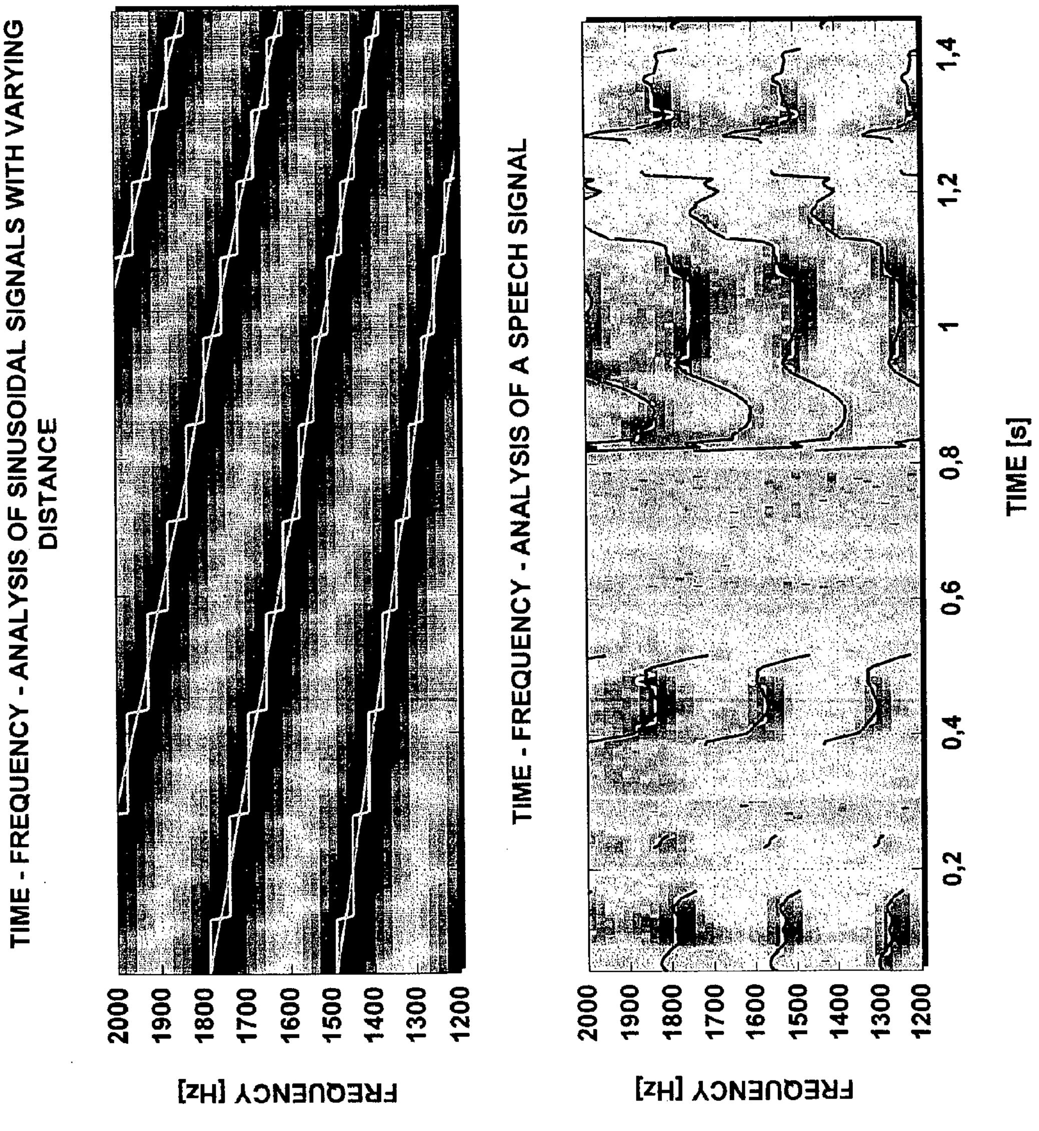

TIME - FREQUENCY - ANALYSIS OF SINUSOIDAL SIGNALS WITH VARYING DISTANCE
FREQUENCY [Hz]
2000
1900
1800
1700
1600
1500
1400
1300
1200
TIME - FREQUENCY - ANALYSIS OF A SPEECH SIGNAL
FREQUENCY [Hz]
2000
1900
1800
1700
1600
1500
1400
1300
1200
0,2
0,4
0,6
0,8
1
1,2
1,4
TIME [s]

FIG. 16

European Patent Office

EUROPEAN SEARCH REPORT

Application Number
EP 07 00 0568

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 Y	QUAST H ET AL: "Robust pitch tracking in the car environment" 2002 IEEE INTERNATIONAL CONFERENCE ON ACOUSTICS, SPEECH, AND SIGNAL PROCESSING. PROCEEDINGS (CAT. NO.02CH37334) IEEE PISCATAWAY, NJ, USA, vol. 1, 2002, pages 353-356, XP002434822 ISBN: 0-7803-7402-9 * page 353, right-hand column, paragraph 4 - page 354, left-hand column, paragraph 2 * -----	1-42 43-51	INV. G10L11/04
A	ROSS M J ET AL: "Average magnitude difference function pitch extractor" IEEE TRANSACTIONS ON ACOUSTICS, SPEECH AND SIGNAL PROCESSING USA, vol. ASSP-22, no. 5, October 1974 (1974-10), pages 353-362, XP002434823 ISSN: 0096-3518 * page 353, right-hand column, paragraph 1 - page 355, left-hand column, paragraph 2 * -----	1-42	TECHNICAL FIELDS SEARCHED (IPC)
A	WO 02/07363 A (IBM [US]; CHAZAN DAN [IL]; ZIBULSKI MEIR [IL]; HOORY RON [IL]) 24 January 2002 (2002-01-24) * page 14, line 21 - page 15, line 7 * -----	1-42	G10L
A	ATKINSON I A ET AL: "Pitch detection of speech signals using segmented autocorrelation" ELECTRONICS LETTERS, IEE STEVENAGE, GB, vol. 31, no. 7, 30 March 1995 (1995-03-30), pages 533-535, XP006002624 ISSN: 0013-5194 * the whole document * ----- -/--	1-42	

The present search report has been drawn up for all claims

Place of search	Date of completion of the search	Examiner
The Hague	10 August 2007	Burchett, Stefanie

CATEGORY OF CITED DOCUMENTS

X : particularly relevant if taken alone
Y : particularly relevant if combined with another document of the same category
A : technological background
O : non-written disclosure
P : intermediate document

T : theory or principle underlying the invention
E : earlier patent document, but published on, or after the filing date
D : document cited in the application
L : document cited for other reasons

& : member of the same patent family, corresponding document

2

European Patent Office

EUROPEAN SEARCH REPORT

Application Number
EP 07 00 0568

DOCUMENTS CONSIDERED TO BE RELEVANT

Category	Citation of document with indication, where appropriate, of relevant passages	Relevant to claim	CLASSIFICATION OF THE APPLICATION (IPC)
Y	KLAPURI A P: "Multiple fundamental frequency estimation based on harmonicity and spectral smoothness" IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING, IEEE SERVICE CENTER, NEW YORK, NY, US, vol. 11, no. 6, November 2003 (2003-11), pages 804-816, XP011104552 ISSN: 1063-6676 * page 806, left-hand column, line 1 - right-hand column, line 36 * -----	43-51	
Y	XU JINFU ET AL: "Noise-robust speech recognition based on difference of power spectrum" ELECTRONICS LETTERS, IEE STEVENAGE, GB, vol. 36, no. 14, 6 July 2000 (2000-07-06), pages 1247-1248, XP006015408 ISSN: 0013-5194 * page 1247, left-hand column, paragraph 7 - right-hand column, paragraph 2 * -----	43-51	TECHNICAL FIELDS SEARCHED (IPC)
A	SHIMAMURA T ET AL: "Noise-robust fundamental frequency extraction method based on exponentiated band-limited amplitude spectrum" CIRCUITS AND SYSTEMS, 2004. MWSCAS '04. THE 2004 47TH MIDWEST SYMPOSIUM ON HIROSHIMA, JAPAN JULY 25-28, 2004, PISCATAWAY, NJ, USA,IEEE, vol. 2, 25 July 2004 (2004-07-25), pages II141-II144, XP010738725 ISBN: 0-7803-8346-X * the whole document * ----- -/--	43-51	

The present search report has been drawn up for all claims

Place of search	Date of completion of the search	Examiner
The Hague	10 August 2007	Burchett, Stefanie

CATEGORY OF CITED DOCUMENTS

X : particularly relevant if taken alone
Y : particularly relevant if combined with another document of the same category
A : technological background
O : non-written disclosure
P : intermediate document

T : theory or principle underlying the invention
E : earlier patent document, but published on, or after the filing date
D : document cited in the application
L : document cited for other reasons

& : member of the same patent family, corresponding document

European Patent Office

EUROPEAN SEARCH REPORT

Application Number
EP 07 00 0568

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	COSI P ET AL: "Auditory modeling techniques for robust pitch extraction and noise reduction" ICSLP 98 : 5TH INTERNATIONAL CONFERENCE ON SPOKEN LANGUAGE PROCESSING.(INCORPORATING 7TH AUSTRALIAN INTERNATIONAL SPEECH SCIENCE AND TECHNOLOGY CONFERENCE). SYDNEY, AUSTRALIA, NOV. 30 - DEC. 4, 1998, INTERNATIONAL CONFERENCE ON SPOKEN LANGUAGE PROCES, vol. CD-ROM, 30 November 1998 (1998-11-30), pages 1053-1057, XP002175877 ISBN: 1-876346-17-5 * page 1054, right-hand column, paragraph 3 - page 1055, left-hand column, paragraph 2 * -----	43-51	
			TECHNICAL FIELDS SEARCHED (IPC)
	The present search report has been drawn up for all claims		

Place of search	Date of completion of the search	Examiner
The Hague	10 August 2007	Burchett, Stefanie

CATEGORY OF CITED DOCUMENTS

X : particularly relevant if taken alone
Y : particularly relevant if combined with another document of the same category
A : technological background
O : non-written disclosure
P : intermediate document

T : theory or principle underlying the invention
E : earlier patent document, but published on, or after the filing date
D : document cited in the application
L : document cited for other reasons

& : member of the same patent family, corresponding document

European Patent
Office

Application Number
EP 07 00 0568

CLAIMS INCURRING FEES

The present European patent application comprised at the time of filing more than ten claims.

□ Only part of the claims have been paid within the prescribed time limit. The present European search report has been drawn up for the first ten claims and for those claims for which claims fees have been paid, namely claim(s):

□ No claims fees have been paid within the prescribed time limit. The present European search report has been drawn up for the first ten claims.

LACK OF UNITY OF INVENTION

The Search Division considers that the present European patent application does not comply with the requirements of unity of invention and relates to several inventions or groups of inventions, namely:

see sheet B

☒ All further search fees have been paid within the fixed time limit. The present European search report has been drawn up for all claims.

□ As all searchable claims could be searched without effort justifying an additional fee, the Search Division did not invite payment of any additional fee.

□ Only part of the further search fees have been paid within the fixed time limit. The present European search report has been drawn up for those parts of the European patent application which relate to the inventions in respect of which search fees have been paid, namely claims:

□ None of the further search fees have been paid within the fixed time limit. The present European search report has been drawn up for those parts of the European patent application which relate to the invention first mentioned in the claims, namely claims:

European Patent
Office

LACK OF UNITY OF INVENTION SHEET B

Application Number
EP 07 00 0568

The Search Division considers that the present European patent application does not comply with the requirements of unity of invention and relates to several inventions or groups of inventions, namely:

1. claims: 1-42

Fundamental frequency estimation of low voices

---

2. claims: 43--51

Fundamental frequency estimation with increased accuracy for speech signals containing background noise

---

ANNEX TO THE EUROPEAN SEARCH REPORT ON EUROPEAN PATENT APPLICATION NO. EP 07 00 0568

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.

10-08-2007

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
WO 0207363 A	24-01-2002	AU 7272901 A	30-01-2002
		CA 2413138 A1	24-01-2002
		CN 1527994 A	08-09-2004
		EP 1309964 A2	14-05-2003
		US 6587816 B1	01-07-2003

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.

Patent documents cited in the description

- EP 06024940 A **[0074] [0074] [0077]**

Non-patent literature cited in the description

- **K. FELLBAUM.** Sprachverarbeitung und Sprachübertragung. Springer, 1984 **[0003] [0011]**
- **D.K. FREEMAN ; G. COSIER ; C.B. SOUTHCOTT ; I. BOYD.** The Voice Activity Detector for the PAN-European Digital Cellular Mobile Telephone Service. *Proceed. of the Intern. Conf. on Acoust., Speech, and Signal Process.,* 1989, vol. 1, 369-372 **[0003]**
- **W. HESS.** Pitch Determination of Speech Signals. Springer, 1983 **[0003] [0010] [0011]**
- **P. VARY ; R. MARTIN.** Digital Speech Transmission. John Wiley & Sons, 2006 **[0003] [0009] [0056] [0086]**
- **P. VARY ; U. HEUTE ; W. HESS.** Digitale Sprachsignalverarbeitung. Teubner, 1998 **[0003] [0009]**
- **E. HÄNSLER ; G. SCHMIDT.** Acoustic Echo and Noise Control - A Practical Approach. John Wiley & Sons, 2004 **[0009] [0056]**
- **E. HÄNSLER ; G. SCHMIDT.** Topics in Acoustic Echo and Noise Control - Selected Methods for the Cancellation of Acoustic Echoes, the Reduction of Background Noise, and Speech Processing. Springer, 2006 **[0009]**
- **J. DELLER ; J. HANSEN ; J. PROAKIS.** Discrete-Time Processing of Speech Signals. IEEE-Press, 1993 **[0010]**
- **M. R. SCHROEDER.** Period Histogram and Product Spectrum: New Methods for Fundamental Frequency Measurements. *J. Acoust. Soc. Am.,* 1968, vol. 43 (4), 829-834 **[0010]**
- **D.K. FREEMAN ; G. COSIER ; C.B. SOUTHCOTT ; I. BOYD.** The Voice Activity Detector for the PAN-European Digital Cellular Mobile Telephone Service, Proceed. of the Intern. Conf. on Acoust. *Speech, and Signal Process.,* 1989, vol. 1, 369-372 **[0011]**
- **J. BENESTY ; S. MAKINO ; J. CHEN.** Speech Enhancement. Springer, 2005 **[0056]**
- Topics in Acoustic Echo and Noise Control - Selected Methods for the Cancellation of Acoustic Echoes. **E. HÄNSLER ; G. SCHMIDT.** Reduction of Background Noise, and Speech Processing. Springer, 2006 **[0056]**
- **S. F. BOLL.** Suppression of Acoustic Noise in Speech Using Spectral Subtraction. *IEEE Trans. Acoust. Speech Signal Process,* 1979, vol. 27 (2), 113-120 **[0086]**
- **E. HÄNSLER.** Statistische Signale - Grundlagen und Anwendungen. Springer, 2001 **[0086]**
- **T. HAULICK ; K. LINHARD.** Noise Subtraction with Parametric Recursive Gain Curves. *Proceed. of the European Conf. on Speech Communications and Technology,* 1999, vol. 6, 2611-2614 **[0086]**