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

(11) EP 1 339 045 A1

(12) EUROPEAN PATENT APPLICATION

(43) Date of publication:
27.08.2003 Bulletin 2003/35

(51) Int Cl.[7]: **G10L 21/02**

(21) Application number: **02004143.0**

(22) Date of filing: **25.02.2002**

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

(71) Applicant: **Sony International (Europe) GmbH**
10785 Berlin (DE)

(72) Inventors:
- **Kemp, Thomas, c/o Advanced Tech. Cr. Stuttgart**
70327 Stuttgart (DE)
- **Tato, Raquel, c/o Advanced Tech. Cr. Stuttgart**
70327 Stuttgart (DE)

(74) Representative: **Müller - Hoffmann & Partner Patentanwälte,**
Innere Wiener Strasse 17
81667 München (DE)

(54) Method for pre-processing speech

(57) A method for pre-processing speech in particular for a method for recognizing speech is suggested in which likelihoods (Lj, LM_{total}, L) for occurrences for speech elements (P1, ..., Pm) based on acoustic feature data (AFD) is suggested, wherein before deriving said likelihoods (Lj, LM_{total}, L) parts of said acoustic feature data (AFD) being representative for frequency bands (F1, ..., FN) which are assumed to be distorted by an additive and band-limited noise signal (NS) are exchanged by exchange feature data (EFD) so as to generate modified acoustic feature data (MAFD), said exchange feature data (EFD) being representative for undisturbed and/or average speech.

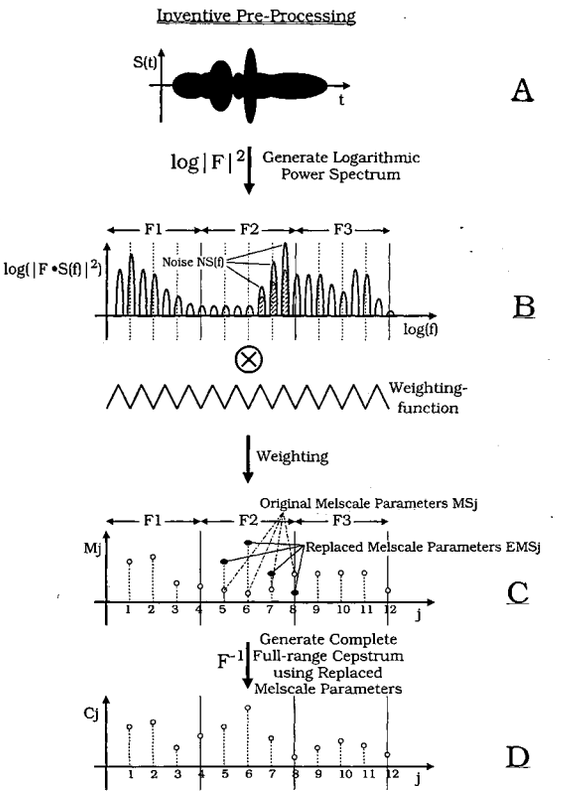

Fig. 3

EP 1 339 045 A1

Description

[0001] The present invention relates to a method for pre-processing speech and in particular to a method for pre-processing speech to be employed in a method for recognizing speech. More particular the invention relates to a method for pre-processing speech using noise-robust acoustic modeling by Union models with back-off.

[0002] Nowadays electronic equipment and appliances become more and more important which use for instance with a man-machine-interface speech dialog systems between a user and the device. Therefore, speech recognition is an important aspect of such appliances. As the distinct applications may also take place with noisy environments the problem of speech recognition from a speech signal which is disturbed or distorted by a noise signal has to be faced.

[0003] Although, many approaches try to separate the noise signal from the speech signal in advance of the recognition process, these known methods are far beyond from leading to a satisfactory recognition rate when noise signals interfere with speech signals.

[0004] Therefore, it is an object of the present invention to provide a method for pre-processing speech, in particular within a method of recognizing speech, which is capable of taking into account noise signals in superposition with speech signals in a particularly simple and reliable manner.

[0005] The object is achieved by a method for pre-processing speech according to the features of claim 1. Additionally, the object is achieved by an apparatus and a computer program product according to the features of claims 19 and 20, respectively. Preferred embodiments of the present invention are within the scope of the dependent subclaims.

[0006] The inventive method for pre-processing speech, in particular in a method for recognizing speech, comprises the steps of receiving a speech signal, analyzing said speech signal with respect to a given number of predetermined frequency bands and thereby generating acoustic feature data which are at least in part representative for said speech signal with respect to said frequency bands. Further the inventive method comprises the step of deriving likelihoods for occurrences of speech elements or of sequences thereof within said speech signal based on said acoustic feature data or a derivative thereof, wherein before deriving said likelihoods parts of said acoustical feature data being representative for frequency bands which are at least assumed to be disturbed or distorted by an additive and band-limited noise signal are exchanged by exchange feature data, so as to generate modified acoustic feature data, said exchange feature data being representative for undisturbed and/or average speech.

[0007] It is therefore a basic idea of the present invention to exchange information extracted from the speech signal which is assumed to be disturbed or distorted by a noise signal by undisturbed and undistorted information. Consequently, the whole content of information upon which the recognition process is based, i.e. on which at least the generation and calculation of likelihoods of speech elements is based, is modified, so as to exchange disturbed or distorted information by artificial information. It is therefore assumed, that the content of artificial information leads to a better recognition process and in particular to a more reliable calculation of likelihoods.

[0008] According to a preferred embodiment of the inventive method of pre-processing speech a set of frequency domain parameters is generated at least as a part of said acoustic feature data. Further, subsets of said frequency domain parameters are assigned to different frequency bands. According to that measure first of all the time domain signal, i.e. the amplitude of the voice signal as a function of time, is converted into a frequency domain signal, upon which certain frequency domain parameters may be obtained being representative for or describing the speech signal in the time domain and/or in the frequency domain. The frequency domain parameters are generated and designed so as to be representative for different frequency bands of the speech signal in the frequency domain. In principle, both the time and the frequency domain carry exactly the same information.

[0009] In a preferred embodiment of the present invention melscale parameters are used as frequency domain parameters. Therefore, the complete frequency range of the received speech signal may be subdivided in and/or covered by a set of frequency bands taking into account the different information contents of the frequency bands or frequency intervals with respect to human perception capabilities.

[0010] According to a further preferred embodiment of the present invention said exchange feature data are chosen to include exchange frequency domain parameters and in particular they may contain or include exchange melscale parameters to exchange frequency domain parameters and in particular melscale parameters which belong or which are assumed to belong to disturbed or distorted frequency bands. Thereby, a modified set of frequency domain parameters is generated.

[0011] It is of particular advantage to derive at least a part of said exchange feature data and in particular at least a part of said exchange frequency domain parameters based on a first acoustic model set which operates on the frequency domain and/or on the space of said frequency domain parameters.

[0012] It is further preferred to use frequency domain parameters and in particular melscale parameters in the process of deriving likelihoods for the occurrence of speech elements within the received speech signal.

[0013] It is a further aspect of the present invention to use a first acoustic model set which is based on the entire frequency range of the speech signal. Alternatively, said first model set may comprise submodels which

base solely on respective frequency bands. In any case, it may be advantageous to involve information from as much frequency bands as possible.

[0014] A further aspect of the present invention is to use as a first acoustic model set an acoustic model set which is based on average speech and/or on undisturbed or undistorted speech. According to that particular measure said derived exchange feature data are free from disturbance compared to the received speech signal, thereby ensuring a better recognition and an higher recognition rate.

[0015] According to an alternative it is also possible to use time domain parameters or time domains like parameters and in particular cepstral parameters for the generation of said likelihoods.

[0016] According to another aspect of the present invention it is suggested to derive said likelihoods using a union-model-like strategy, wherein it is assumed that a given and fixed number of a frequency bands or frequency intervals which is lower than the number of the entirety of the frequency bands are disturbed or distorted by a band-limited and additive noise signal.

[0017] It is further preferred that for each assumed disturbed or distorted frequency band the corresponding frequency domain parameters and in particular the corresponding melscale parameters are exchanged by exchange frequency domain parameters which correspond to a speech element P1, ..., Pm to be tested and in particular taken from said first acoustic model to generate modified acoustic feature data.

[0018] It is further provided in the inventive method to derive based on said frequency domain parameters and said exchange frequency domain parameters said time domain parameters or time domain like parameters and in particular said cepstral parameters, in particular by involving an inverse Fourier transform or the like. Although, it follows from actual calculation schemes that e. g. cepstral parameters are not true time domain parameters, they may be referred to as time domain like parameters as they are derived by involving an inverse Fourier transform leading back from the frequency domain.

[0019] According to a further advantageous embodiment of the present invention the likelihoods are derived based on said time domain parameters, time domain like parameters and/or in particular they are based on said cepstral parameters.

[0020] It is a further aspect to generate single likelihoods for each combination of assumed and fixed M distorted frequency bands. Based on these single likelihoods for the variety of M assumed distorted frequency bands a total likelihood is derived.

[0021] According to a further aspect these total likelihoods are derived for each number M of assumed distorted frequency bands, said number M satisfying the relation $0 \leq M < N$. I.e., it is assumed that the number of assumed distorted frequency bands is lower than the number of the entirety of frequency bands of the speech signal. The case M = 0 represents the case without noise. Under these circumstances a global likelihood is derived from said global likelihoods for each M and for each combination of assumed distorted or disturbed frequency bands.

[0022] These both latter aspects together with the inventive principle of exchanging corrupted frequency domain data represent a modification of the union-model-like strategy, which is also a basic aspect of the present invention. According to these aspects the global likelihood for the occurrence of a speech element is calculated on the entirety of single likelihoods for each combination and for each number of assumed distorted frequency bands.

[0023] In the following the construction and generation of the single likelihoods Lj, the total likelihoods LM_{total} and the global likelihood L a decomposition of the whole frequency range F of a received speech input into three frequency subbands F1, F2, and F3 is assumed. It is further assumed that the occurance of a particular speech element X in said received speech input is tested.

[0024] Further, in the following Lj is representative for the likelihood of the occurance of speech element X in said speech input based on frequency range or subband Fj. Therefore, single likelihoods L1, L2, and L3 are calculated.

[0025] For each assumption of $M < N = 3$ distorted frequency bands a total likelihood $L1_{total}$, $L2_{total}$ can be derived in an approximative way:

$$\text{for } M = 1 : L1_{total} \approx L1 \cdot L2 + L1 \cdot L3 + L2 \cdot L3$$

$$\text{for } M = 2 : L2_{total} \approx L1 + L2 + L3.$$

[0026] For a simple estimation the global likelihood L can be set to

$$L := L1_{total} \text{ or } L := L2_{total}.$$

[0027] If, however, the number of distorted subbands is unknown all terms from the equation listed above can be added to contribute to the global likelihood L:

$$L := L1 \cdot L2 + L1 \cdot L3 + L2 \cdot L3 + L1 + L2 + L3$$

$$:= L1_{total} + L2_{total}.$$

[0028] It is a further aspect of the present invention, that this calculation scheme can be used to describe the likelihood of the occurance of a speech element X in the received speech input. This scheme is based on the fact that with a proper and appropriate replacement by modified acoustic feature data the approximation

$$L1 \cdot L2 \approx L1$$

holds.

[0029] It is of further advantage to use for deriving said likelihoods a second acoustic model which operates on the time domain or time-like domain and in particular on the space of the time domain parameters time domain like parameters and/or more particular on the space of the cepstral parameters. Additionally, in deriving said likelihoods, the complete cepstral information is used according to the invention.

[0030] To increase the performance of the inventive method it is suggested to use a first acoustic model in the frequency domain having different complexities compared to said second acoustic model operating in the time domain.

[0031] According to a further aspect of the invention a system, an apparatus, a device, a dialog system, and/or the like is provided which is in each case adapted to realize, to carry out and/or to perform a method for pre-processing speech according to the present invention and/or the steps thereof.

[0032] It is a further aspect of the present invention to provide a computer program product, comprising computer program means which is adapted to perform and/or to realize the inventive method for pre-processing speech and/or the steps thereof, when it is executed on a computer, a digital signal processing means, and/or the like.

[0033] It has to be emphasized, that according to the preferred embodiment of the present invention two acoustic models are involved in the inventive method. The first acoustic model operates in the frequency domain and is used to exchange distorted information by undistorted information. Based on the thereby modified acoustic feature data information in the time domain is obtained. The likelihoods for different speech elements are extracted based on said time domain information using the second acoustic model which operates in the time domain or time-like domain.

[0034] These and further aspects of the present invention are discussed taking reference to the following remarks:

Speech recognition in the presence of noise is a difficult problem of great practical importance. Recently, the Union model has been proposed that tries to overcome the signal quality deterioration by the assumption of band-limited additive noise, and by effectively ignoring the contribution of the distorted signal band in the likelihood computation.

[0035] However, this model suffers from the reduction of the information usable by the likelihood computation.

[0036] The present invention overcomes this limitation and makes the full information usable by replacing the corrupted band by an estimate of the average speech information inside that band.

[0037] To achieve this object an improved Union model for the likelihood combination in the presence of band-limited additive noise is proposed.

Basically, a signal is split up in N (let N = 5 for clarity from now on, however N is arbitrary) frequency bands. Under the assumption that M (M < N) bands are distorted (let M = 1 for clarity from now on although the algorithm is not limited to M = 1, the likelihood for a speech element, e.g. a phoneme, can be computed as the sum of the likelihood contributions of all combinations of N - M (= 4) bands.

[0038] The principal idea is that if a combination includes the corrupted, i. e. noisy, band then its likelihood is very low, and therefore the sum of the individual likelihood contributions is dominated by the one combination of bands where the noisy bands is excluded.

[0039] The interesting property of the Union model is now that it is not necessary to know which of the bands is corrupted.

[0040] In the Union model, in order to run the pre-processing on the individual subbands, it is no longer possible, to compute features that use input from different sub-bands. Specifically, in the Union model it is only possible to compute a limited number, e. g. in a specific implementation 3 cepstral parameters, which can be associated with C4, C5 and C6 in the standard model. The parameters C1, C2 and C3 carry important information, this is a strong drawback of the standard Union model as such which leads to a reduced performance of the Union model if there is no noise present.

[0041] In the presence of noise, the Union model suffers less than the standard model does. However, in many cases the increased robustness is not sufficient to compensate the lack of performance of the baseline, i. e. no noise, case.

[0042] It is a basic idea of the invention to allow for the full number of cepstral parameters to be computed by changing the pre-processing significantly, while maintaining the principle advantage of the Union model, i. e. the independence of the identity of the corrupted channel.

[0043] The reason for the intrinsic inability of the Union model to make use of less cepstra, and no lower order cepstra is its very idea, the sub-band processing. If the full range of, say, Q (typically: 15 < Q < 40) spectral energy parameters is split up in N bands of size Q/N each, then the discrete cosine transform that is used to decorrelate the features can only be computed up to order (Q/2N), rather than to order (Q/2) as in the original model. In order to compute low-or-der cepstra, it is necessary to take into account information from all bands, which is not feasible in the Union.

[0044] It is suggested to have two sets of acoustic models, i. e. the standard models P in the cepstral domain - without any sub-streams, using the whole information - and another set of models K in the log spectral domain.

[0045] The models K can be of different complexity. Namely, they can have one mean vector per HMM-

state-model of the recognizer, e. g. for a monophone recognizer, one mean vector per phoneme. Or they can have only one global mean vector for all speech and silence, or one for speech and one for silence, or any other degree of tying between this two extremes.

[0046] Suppose now in the union model the number of streams or frequency bands is 3, and the number of corrupted streams is 1. The number of log spectral coefficients totals e. g. 21. Then, in the standard union model the final likelihood is computed as (L1*L2) + (L1*L3) + (L2*L3), where e. g. L2 is the likelihood computed only on the second stream or band, using some cepstral parameters derived from the mid part of the spectrum or coefficients 8-14 of the 21 log spectral coefficients.

[0047] In this invention, it is suggested to compute the terms, e. g. comparable to L1* L3, differently. Rather than constructing models for streams L1 and L3, evaluating them separately, and multiplying the likelihood, it is proposed to use the original cepstral models P which are based on all 21 log spectral coefficients, but to replace before deriving the cepstrum the coefficients 8-14 of the log-spectral domain parameters by the coefficients 8-14 of the model K mentioned above. After the replacement has been done, the pre-processing continued as usual, i.e. the cepstral parameters are computed and evaluated using the standard models.

[0048] To compute the term comparable to L1*L2, similarly, the standard pre-processing of the input data is done until the point where the 21 log spectral coefficients are computed. Then, the coefficients 15-21 are replaced by the value taken from the model K, and the pre-processing is continued to compute ordinary cepstral parameters as usual.

[0049] A basic idea is the following. If, say band 3 - corresponding to the coefficients 15-21 in log spectral domain - is corrupted, then it conveys no useful information any more. It should therefore not be used to compute the likelihoods. However, we can "reconstruct" the corrupted information, by replacing it with the average information that is contained in this band as averaged over all the speech, or by the average information usually found in this band for this phone. Clearly, since the artificial data inserted into band 3 is the same for all phone models being evaluated, it cannot discriminate between them and adds no information whatsoever to the discriminative power of the model. But, it allows the computation of lower order cepstra which relate all other bands with each other.

[0050] So, the net effect of the replacement is to blur information in band 3 up to the point where it is unusable, but to keep the information in bands 1 and 2 available, and also the information about the relationship of bands 1 and 2 available which is not available in the union model. Since the lower order cepstral parameters contain exactly this type of information, they cannot be used in the union model, but they can be computed in the proposed model.

[0051] No information loss is incurred in addition to the information loss by the noise, which is theoretically unavoidable. This recovery of lost information is the main advantageous difference between the invention and the state of the art.

[0052] The aspects of the present invention will be discussed in further detail taking reference to the accompanying drawings.

Fig. 1 shows a pre-processing sequence known in the art.

Fig. 2 shows the standard Union model processing in addition to the processing of Fig. 2.

Fig. 3 shows a preferred embodiment of the inventive method for pre-processing speech.

[0053] In the following same reference symbols refer to comparable elements and aspects.

[0054] First of all the standard pre-processing of a received speech signal S is described taking reference to sections A to D of Fig. 1.

[0055] Section A of Fig. 1 shows the envelope of a speech signal S as a function of time S(t).

[0056] Based on this particular amplitude time relationship of the speech signal S the logarithmic power spectrum is generated. This is done by first applying a Fourier transform to the speech signal S. Then the logarithm of the absolute square of the Fourier transformed signal is generated.

[0057] The result of the generation of the logarithmic power spectrum is shown in section B of Fig. 1, wherein for simplicity also the logarithm log(f) of the frequency f is taken. The whole frequency range F is built up by a union of three distinct frequency bands F1, F2, and F3.

[0058] The frequency bands F1 to F3 are subdivided. For each subdivision the average of the logarithmic power spectrum is taken taking into account a weighting function, which is in the case of section B of Fig. 1 a piecewise triangular weighting function.

[0059] The result of piecewise averaging the logarithmic power spectrum with the triangular weighting function is shown in section C of Fig. 1, where on the ordinate the average value Mj of the distinct subdivisions numbered by j are shown as single values. According to section C of Fig. 1 to each subdivision with number j a parameter Mj are in the frequency domain is assigned. These parameters Mj called melscale parameters and they are examples of the frequency domain parameters in the sense of the invention.

[0060] Although, the derived melscale parameters Mj may be used as said frequency domain parameters MSJ in the sense of the invention and in particular for generating the likelihoods of the distinct speech elements P1, ..., Pm within the speech signal S, it is often more appropriate to generate from the melscale parameters Mj the so-called cepstral parameters Cj which built up

the cepstrum corresponding to the spectrum. This is done by essentially applying a discrete inverse Fourier transform to the set of melscale parameters Mj of section C. The result is shown in section D of Fig. 1.

[0061] The so derived cepstral parameters Cj are used as input values for an acoustic model P operating in the time domain, time-like domain or the domain of the cepstral parameters C1. Upon input of the cepstral parameters Cj and an speech element X to be tested a likelihood is obtained being descriptive for the chance of occurrences of the tested speech element X within the speech signal S(t) or a part thereof.

[0062] It is emphasized that the distinct amplitudes of the signals and the derived parameters of sections A to D of Fig. 1 have no strict mathematical correspondence to each other but are only used as a simplified explanation for the relationship between them.

[0063] In the union model approach of the prior art shown in Fig. 2 the pre-processing scheme differs with respect to the transition from the section C to section D of the processing of Fig. 1.

[0064] After calculating the melscale parameters MSj as frequency domain parameters according to section C of Fig. 2 the melscale domain or frequency domain is subdivided with respect to the given and predetermined frequency bands F1 to F3.

[0065] The result is shown in section E of Fig. 2. For each frequency band F1 to F3 a separate set of melscale parameters M1 - M4, M5 - M8 and M9 - M10 are derived. In contrast to the processing shown in Fig. 1, where the entirety of all melscale parameters M1 - M12 is used to calculate the cepstral parameters C1 - C12 by applying an inverse Fourier transform, cepstral parameters are calculated for each of the subdivided melscale domains separately.

[0066] That means, that by applying an inverse Fourier transform to the set M1 - M4 cepstral parameters $C1^1$ - $C4^1$ are derived. Because of the properties of the inverse Fourier transform and because of the subdivision and separation of the melscale domain these first parameter set $C1^1$ - $C4^1$ does not contain information from the other frequency bands F2 and F3. The parameters $C1^1$ - $C4^1$ are at most comparable to the cepstral parameters C5 - C8 of the processing of Fig. 1, the latter of which taking into account the whole frequency information of all frequency bands F1 - F3.

[0067] As a result, in the processing of Fig. 2 information is lost with respect to the lower cepstral parameters C1 to C4 of Fig. 1, although the aspect of noise is taken into account by the union model processing of Fig. 2.

[0068] Fig. 3 shows an embodiment of the inventive pre-processing of a speech signal S.

[0069] In section B of Fig. 3 it is shown, that in frequency band F2 noise components NS(f) are added. The scattered frequency components show the components which are comparable with section B of Fig. 1, i. e. the noise-free case.

[0070] In the transition from section B to section C of Fig. 3 again a weighting function is applied to the logarithmic power spectrum of section B leading to a piecewise average with respect to the frequency subdivisions of the frequency bands F1 - F3. As shown in section C of Fig. 3 the resulting melscale parameters M5 - M8, i. e. the frequency domain parameters MS5 - MS8 in the sense of the invention, are replaced by exchange melscale parameters EMS5 - EMS8 which are represented by filled symbols and which are taken from an acoustic model K with respect to undisturbed speech speech in the frequency domain.

[0071] The set of melscale parameters of section C, where the original melscale parameters MS5 - MS8 are replaced by corrected melscale parameters EMS5 - EMS8 are used to again generate a complete and full range cepstrum as shown in section D of Fig. 3.

[0072] The major difference between the processing of the prior art according to Fig. 1 and the inventive processing of Fig. 3 is that a noise aspect is taken into account by replacing melscale parameters which are assumed to belong to distorted signal components by parameters which are undisturbed. The difference between the prior art processing of Fig. 2 and the inventive processing of Fig. 3 is that the complete frequency domain information of all sub-bands F1 - F3 is used to calculate the time domain parameters or cepstral parameters Cj.

[0073] The union model strategy is incorporated into the processing of Fig. 3 by taking into account all combinations of likelihoods for distorted frequency bands F1 - F3. As this is done, it is not necessary to know which frequency band actually is distorted by noise.

[0074] Again all numbers of assumed distorted frequency bands can be taken into account as long as the number of assumed distorted frequency bands is lower than the number of frequency bands at all.

Claims

1. Method for pre-processing speech, in particular in a method for recognizing speech, comprising the steps of:

   - receiving a speech signal (S),
   - analyzing said speech signal (S) with respect to a given number (N) of predetermined frequency bands (F1, ..., FN),
   - thereby generating acoustic feature data (AFD) which are at least in part representative for said speech signal with respect to said frequency bands (F1, ..., FN),
   - deriving likelihoods for occurrences of speech elements (P1, ... Pm) or of sequences thereof within said speech signal (S) based on said acoustic feature data (AFD) or a derivative thereof,
   - wherein before deriving said likelihoods parts

of said acoustic feature data (AFD) being representative for frequency bands (F1, ..., FN) which are at least assumed to be disturbed by an additive and band-limited noise signal (NS) are exchanged by exchange feature data (EFD) so as to generate modified acoustic feature data (MAFD), said exchange feature data (EFD) being representative for undisturbed and/or average speech.

2. Method according to claim 2,

- wherein a set of frequency domain parameters (MS1, ..., MS12) is generated as a part of said acoustic feature data (AFD), and
- wherein subsets of said frequency domain parameters (MS1, ..., MS12) are assigned to different frequency bands (F1, ..., FN).

3. Method according to claim 2,
wherein melscale parameters are used as frequency domain parameters (MS1, ..., MS12).

4. Method according to any one of claims 2 or 3,
wherein said exchange feature data (EFD) are chosen to include exchange frequency domain parameters (EMS1, ..., EMS12) and in particular exchange melscale parameters,
to exchange frequency domain parameters (MS1, ..., MS12) and in particular melscale parameters belonging to disturbed frequency bands (F1, ..., FN) to obtain a modified set of frequency domain parameters.

5. Method according to any one of the preceding claims,
wherein at least a part of said exchange feature data (EFD) and in particular said exchange frequency domain parameters (EMS1, ..., EMS12) are derived and/or taken from a first acoustic model (K) operating on the frequency domain and in particular on the space of said frequency domain parameters (MS1, ..., MS12).

6. Method according to claim 5,
wherein a first acoustic model set (K) is used which is based on the entire frequency range (F) of the speech signal (S) or which includes submodels solely based on respective frequency bands (F1, ... FN).

7. Method according to any one of the claims 5 or 6,
wherein a first acoustic model set (K) is used which is based on average and/or undisturbed speech.

8. Method according to any one of the preceding claims,
wherein frequency domain parameters (MS1, ..., MS12) and in particular melscale parameters are used in deriving said likelihoods.

9. Method according to any one of the preceding claims,
wherein time domain parameters or time domain like(C1, ..., C12) are used in deriving said likelihoods.

10. Method according to claim 9,
wherein cepstral parameters (C1, ..., C12) are used as said time domain parameters or said time domain like parameters.

11. Method according to any one of the preceding claims,
wherein said likelihoods are derived using a union-model-like strategy and

- wherein it is assumed that a given and fixed number (M) of frequency bands (F1, ..., FN) lower than said number (N) of the frequency bands (F1, ..., FN) are disturbed or distorted by a band-limited and additive noise signal (NS).

12. Method according to claim 11,
wherein for each assumed disturbed or distorted frequency band (F1, ..., FN) corresponding or assigned frequency domain parameters (MS1, ..., MS12) and in particular corresponding melscale parameters are exchanged by exchange frequency domain parameters (EMS1, ..., EMS12) which correspond to a speech element (P1, ..., Pn) to be tested and in particular taken from said first acoustic model (K) to generate modified acoustic feature data (MAFD).

13. Method according to claim 12,
wherein based on said frequency domain parameters (MS1, ..., MS12) and said exchange frequency domain parameters (EMS1, ..., EMS12) said time domain parameters or time domain like parameters (C1, ..., C12) and in particular said cepstral parameters are derived, in particular by involving an inverse Fourier transform or the like.

14. Method according to any one of the claims 11 to 13,
wherein said likelihoods are derived based on said time domain parameters or time domain like parameters (C1, ..., C12) and in particular based on said cepstral parameters.

15. Method according to any one of the claims 11 to 14,

- wherein single likelihoods (Lj) or substream likelihoods for each combination of fixed M assumed disturbed or distorted frequency bands (F1, ..., FN) are generated and

- wherein a total likelihood (LM_{total}) is derived from said single likelihoods (Lj).

16. Method according to any one of the claims 11 to 15,

- wherein total likelihoods (LM_{total}) are derived for each M fulfilling $0 \leq M < N$ and
- wherein a global likelihood (L) is derived from said total likelihoods (LM_{total}) for each M and for each combination of assumed distorted or disturbed frequency bands (F1, ..., FN).

17. Method according to any one of the preceding claims,
wherein for deriving said likelihoods (Lj, LM_{total}, L) a second acoustic model (P) operating on the time domain or time-like domain and in particular operating on the time domain parameter space, time domain like parameter space and/or the cepstral parameter space is used, evaluating the complete cepstral information.

18. Method according to any one of the preceding claims,
wherein the first acoustic model (K) has different complexities compared to said second acoustic model (P).

19. Apparatus which is capable of realizing a method for pre-processing speech according to any one of the claims 1 to 18 and/or the steps thereof.

20. Computer program product, comprising computer program means adapted to perform and/or to realize a method for pre-processing speech according to any one of the claims 1 to 18 and/or the steps thereof when it is executed on a computer, a digital signal processing means and/or the like.

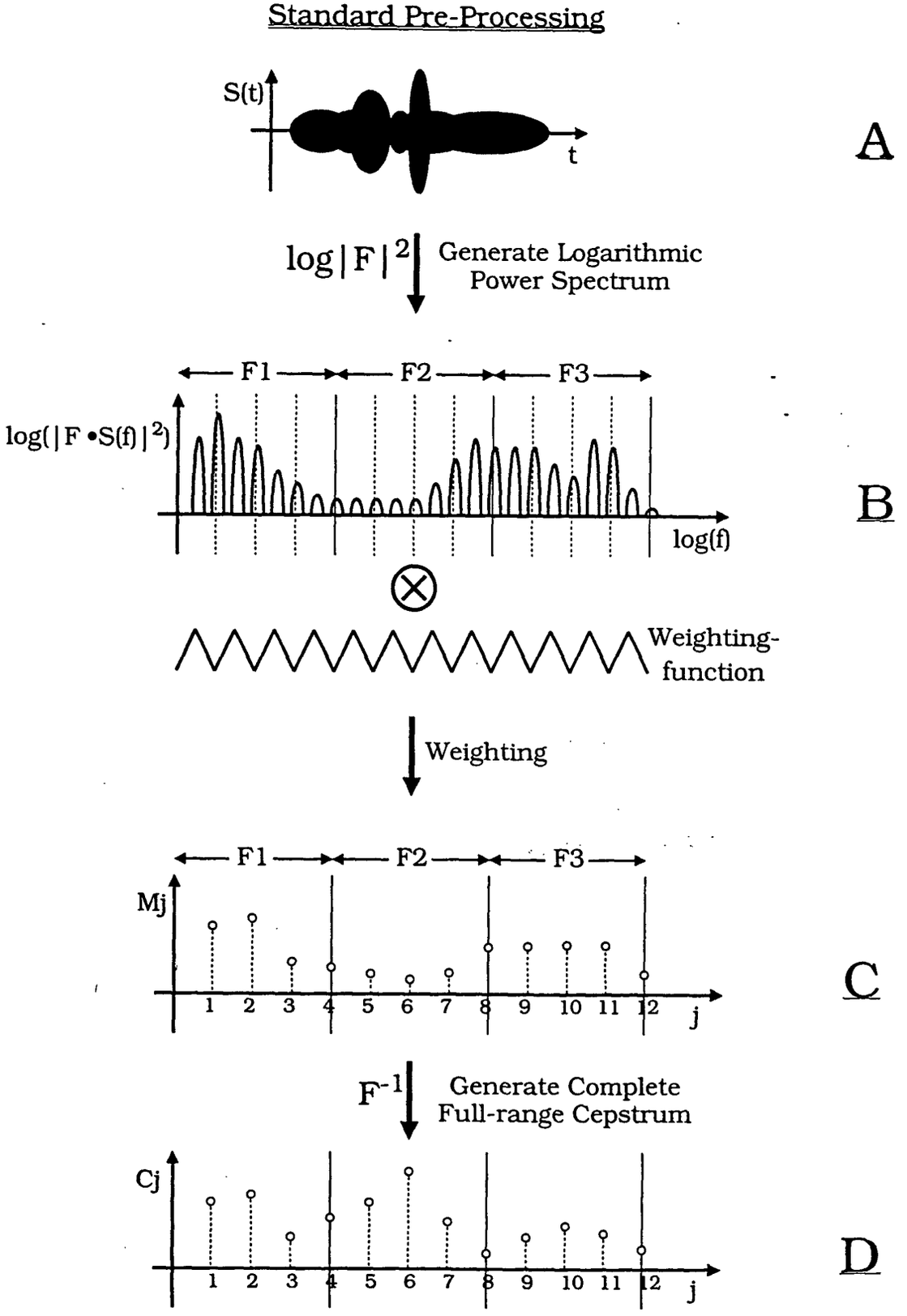
Standard Pre-Processing
S(t)
t
A
log|F|2
Generate Logarithmic Power Spectrum
F1
F2
F3
log(|F •S(f)|2)
log(f)
B
Weighting-function
Weighting
Mj
1 2 3 4 5 6 7 8 9 10 11 12 j
C
F-1
Generate Complete Full-range Cepstrum
Cj
1 2 3 4 5 6 7 8 9 10 11 12 j
D

Fig. 1

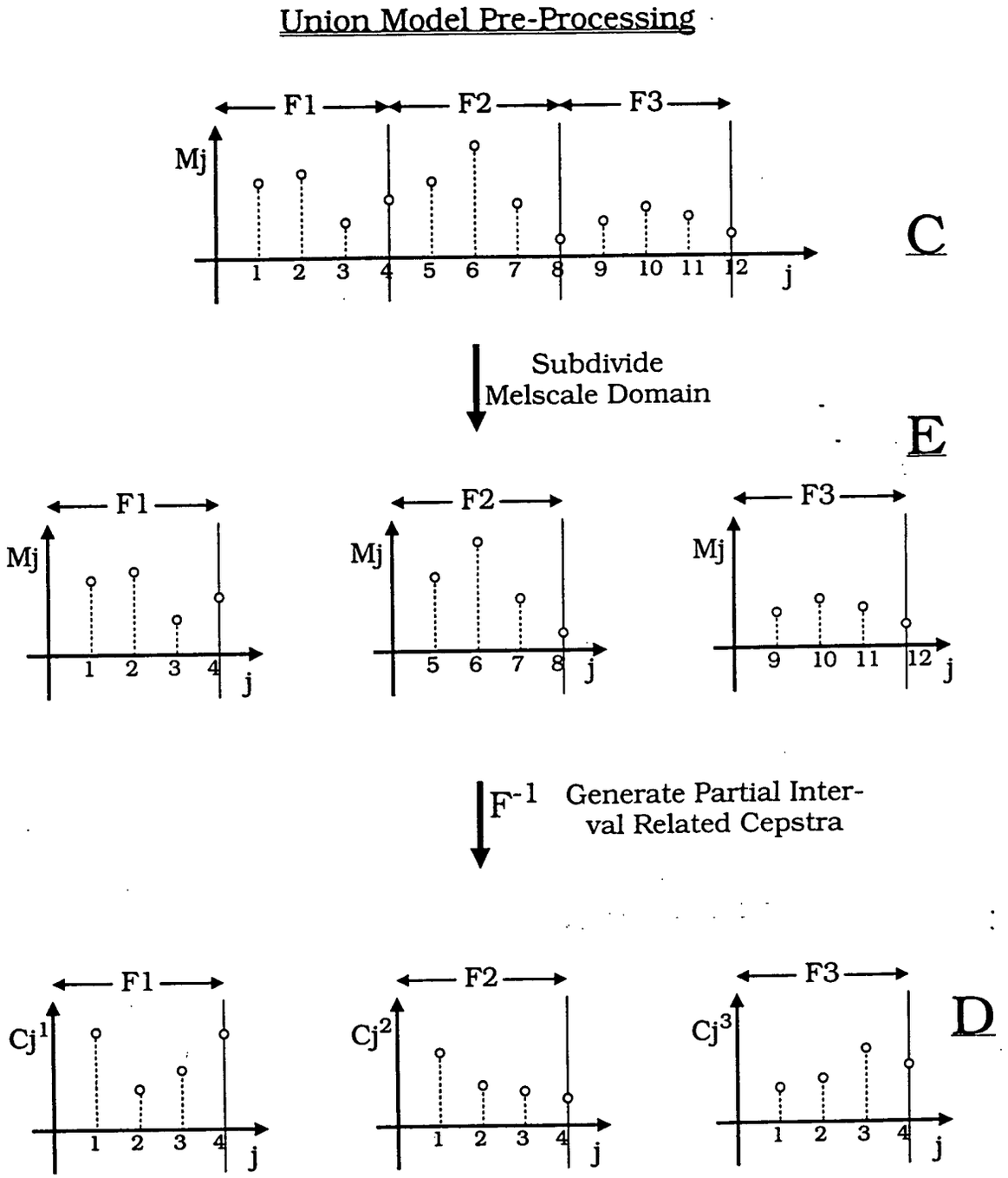
Union Model Pre-Processing
F1
F2
F3
Mj
1 2 3 4 5 6 7 8 9 10 11 12 j
C
Subdivide
Melscale Domain
E
F1
Mj
1 2 3 4 j
F2
Mj
5 6 7 8 j
F3
Mj
9 10 11 12 j
F-1 Generate Partial Interval Related Cepstra
F1
Cj1
1 2 3 4 j
F2
Cj2
1 2 3 4 j
F3
Cj3
1 2 3 4 j
D

Fig. 2

Inventive Pre-Processing

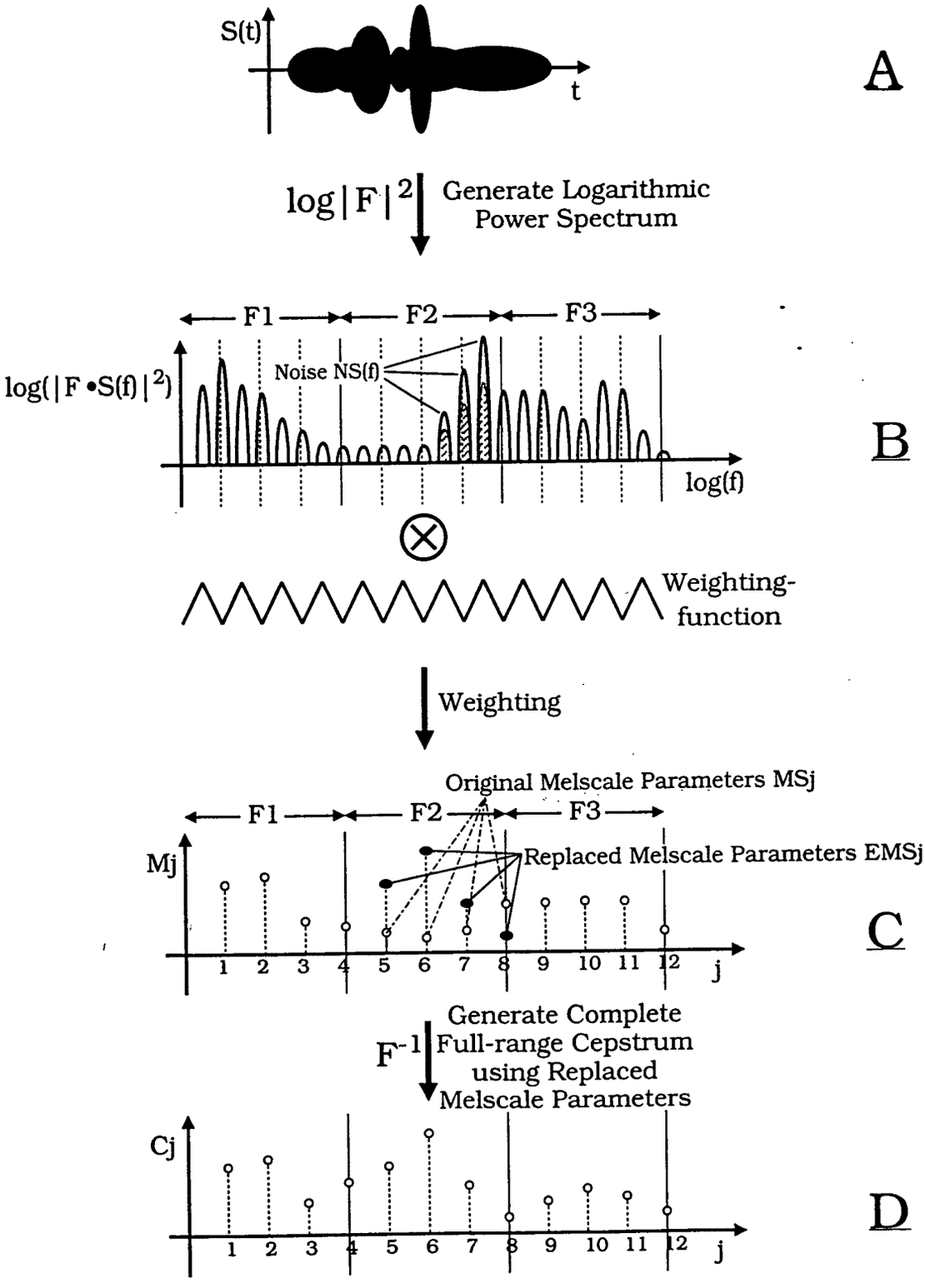

Fig. 3

European Patent Office

EUROPEAN SEARCH REPORT

Application Number
EP 02 00 4143

DOCUMENTS CONSIDERED TO BE RELEVANT

Category	Citation of document with indication, where appropriate, of relevant passages	Relevant to claim	CLASSIFICATION OF THE APPLICATION (Int.Cl.7)
A	US 2001/021905 A1 (NG LAWRENCE C ET AL) 13 September 2001 (2001-09-13) * claim 13 *	1,5,7	G10L21/02
A	EP 0 789 349 A (CANON KK) 13 August 1997 (1997-08-13) * abstract; claims 28-33 *	1,5,7	

TECHNICAL FIELDS SEARCHED (Int.Cl.7)

G10L

The present search report has been drawn up for all claims

Place of search	Date of completion of the search	Examiner
THE HAGUE	20 June 2002	Van Doremalen, J

CATEGORY OF CITED DOCUMENTS

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

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

& : member of the same patent family, corresponding document

EPO FORM 1503 03.82 (P04C01)

ANNEX TO THE EUROPEAN SEARCH REPORT ON EUROPEAN PATENT APPLICATION NO. EP 02 00 4143

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.

20-06-2002

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 2001021905 A1	13-09-2001	US 6377919 B1	23-04-2002
		US 6006175 A	21-12-1999
		AU 3003800 A	04-09-2000
		EP 1163667 A1	19-12-2001
		WO 0049600 A1	24-08-2000
		EP 0883877 A1	16-12-1998
		JP 2000504848 T	18-04-2000
		WO 9729481 A1	14-08-1997
EP 0789349 A	13-08-1997	EP 1213707 A1	12-06-2002
		EP 0789349 A2	13-08-1997
		JP 9244686 A	19-09-1997
		US 2002032566 A1	14-03-2002
		US 5960395 A	28-09-1999

EPO FORM P0459

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