649029

P/00/008b 12/11/91
Section 29 (1)
Regulation 3.1 (2)

AUSTRALIA

Patents Act 1990

NOTICE OF ENTITLEMENT

We, **SRI INTERNATIONAL** of, 333 Ravenswood Avenue, Menlo Park, California 94025, United States of America being the applicant in respect of Application No. 74875/91 state the following:-

The Persons nominated for the grant of the patent have entitlement from the actual inventor by virtue of assignment.

The persons nominated for the grant of the patent is the applicant of the application listed in the declaration under Article 8 of the PCT.

The basic application listed on the request form is the first application made in a Convention country in respect of the invention.

SRI INTERNATIONAL

By our Patent Attorneys,
WATERMARK PATENT & TRADEMARK ATTORNEYS

.....

March 7, 1994

Richard C. T. Smoorenburg
Registered Patent Attorney

AU9174875

(12) PATENT ABRIDGMENT (11) Document No. AU-B-74875/91
(19) AUSTRALIAN PATENT OFFICE (10) Acceptance No. 649029

(54) Title
METHOD FOR SPECTRAL ESTIMATION TO IMPROVE NOISE ROBUSTNESS FOR SPEECH RECOGNITION

(51)[5] International Patent Classification(s)
G10L 003/00

(21) Application No. : 74875/91 (22) Application Date : 25.02.91

(87) PCT Publication Number : WO91/13430

(30) Priority Data

(31) Number	(32) Date	(33) Country
486462	28.02.90	US UNITED STATES OF AMERICA

(43) Publication Date : 18.09.91

(44) Publication Date of Accepted Application : 12.05.94

(71) Applicant(s)
SRI INTERNATIONAL

(72) Inventor(s)
ADORAM ERELL; MITCHEL WEINTRAUB

(74) Attorney or Agent
WATERMARK PATENT & TRADEMARK ATTORNEYS , Locked Bag 5, HAWTHORN VIC 3122

(56) Prior Art Documents
EP 240330

(57) Claim

1. In a speech recognition system for responding to signals representative of digitized speech, a method for preprocessing a noisy speech signal to estimate a clean speech signal, comprising the steps of:

categorizing each time-invariant segment of digitized speech and noise into a noisy speech vector, each element of said noisy speech vector representing one frequency channel of speech information in the presence of noise, said noise being assumed to be additive, time invariant and uncorrelated to the frequency domain;

estimating a clean speech signal comprising a clean speech vector for each said element from said noisy speech vector by

(1) determining the probability distribution of clean speech by applying the noisy speech vector to a mixture model composed of a mixture of components, each component representing a different speech class assuming different frequency channels are uncorrelated within each class,

(2) modelling the conditional probability function of a vector of noisy speech around a vector of clean speech as uncorrelated at different frequency channels; and

storing the estimated clean speech signal in memory.

.../2

10. In a speech recognition system for responding to signals representative of digitized speech, a method for preprocessing speech in the presence of noise to produce an estimated clean speech signal, comprising the steps of:

providing a data base of clean speech;

training a Markov model with said clean speech data base, each state of said Markov model representing a different speech class assuming different frequency channels are uncorrelated within each class, to produce parameters for said Markov model;

categorizing each time-invariant segment of digitized speech and noise into a noisy speech vector, each element of said noisy speech vector representing one frequency channel of speech information in the presence of noise, said noise being assumed to be frequency independent and time invariant;

estimating a clean speech signal comprising a clean speech vector for each vector of a sequence of said noisy speech vectors S'_O to S'_T by

1) determining a probability distribution of the clean speech vector by applying the noisy speech vector to the Markov model using said parameters, and

2) modelling a conditional probability function of a vector of noisy speech around a vector of clean speech as uncorrelated at different frequency channels.

OPI DATE 18/09/91 APPLN. ID 74875 / 91
AOJP DATE 24/10/91 PCT NUMBER PCT/US91/01333
INTERN/ I TREATY (PCT)

(51) International Patent Classification 5 : G10L 3/00

A1

(11) International Publication Number: WO 91/13430

(43) International Publication Date: 5 September 1991 (05.09.91)

(21) International Application Number: PCT/US91/01333

(22) International Filing Date: 25 February 1991 (25.02.91)

(30) Priority data:
486,462 28 February 1990 (28.02.90) US

(71) Applicant: SRI INTERNATIONAL [US/US]; 333 Ravenswood Avenue, Menlo Park, CA 94025-3493 (US).

(72) Inventors: ERELL, Adoram ; Tel Aviv University, Electronic Systems, Faculty of Engineering, 69 978 Ramat Aviv (IL). WEINTRAUB, Mitchel ; 36360 Coronado Drive, Fremont, CA 94536 (US).

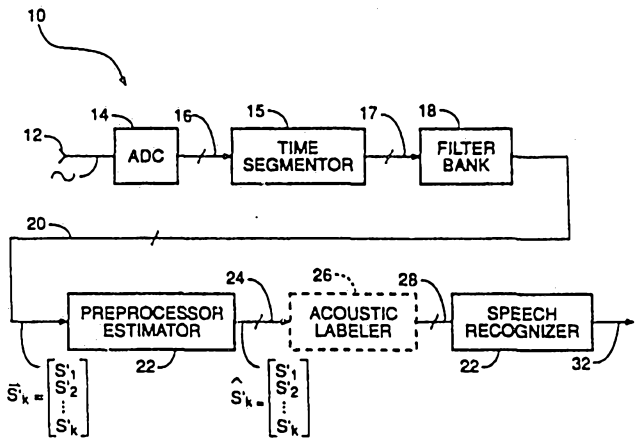
(74) Agents: CHEN, John, Y.; SRI International, 333 Ravenswood Avenue, Menlo Park, CA 94025-3493 (US) et al.

(81) Designated States: AT (European patent), AU, BE (European patent), CA, CH (European patent), DE (European patent), DK (European patent), ES (European patent), FR (European patent), GB (European patent), GR (European patent), IT (European patent), JP, KR, LU (European patent), NL (European patent), SE (European patent).

Published
With international search report.
Before the expiration of the time limit for amending the claims and to be republished in the event of the receipt of amendments.

649029

(54) Title: METHOD FOR SPECTRAL ESTIMATION TO IMPROVE NOISE ROBUSTNESS FOR SPEECH RECOGNITION



(57) Abstract

A method is disclosed for use in preprocessing noisy speech to minimize likelihood of error in estimation for use in a recognizer. The computationally-feasible technique, herein called Minimum-Mean-Log-Spectral-Distance (MMLSD) estimation using mixture models and Markov models, comprises the steps of calculating for each vector of speech in the presence of noise corresponding to a single time frame, an estimate of clean speech, where the basic assumptions of the method of the estimator are that the probability distribution of clean speech can be modeled by a mixture of components each representing a different speech class assuming different frequency channels are uncorrelated within each class and that noise at different frequency channels is uncorrelated. In a further embodiment of the invention, the method comprises the steps of calculating for each sequence of vectors of speech in the presence of noise corresponding to a sequence of time frames, an estimate of clean speech, where the basic assumptions of the method of the estimator are that the probability distribution of clean speech can be modeled by a Markov process assuming different frequency channels are uncorrelated within each state of the Markov process and that noise at different frequency channels is uncorrelated.

METHOD FOR SPECTRAL ESTIMATION TO IMPROVE NOISE ROBUSTNESS FOR SPEECH RECOGNITION

COPYRIGHT NOTICE

A portion of the disclosure of this patent document contains material which is subject to copyright protection. The copyright owner has no objection to the facsimile reproduction by anyone of the patent document or the patent disclosure as it appears in the Patent and Trademark Office patent file or records, but otherwise reserves all copyright rights whatsoever.

BACKGROUND OF THE INVENTION

This invention was supported in part by a grant from the National Science Foundation (IRI-8720403) and in part by SRI International of Menlo Park, California.

This invention relates to speech recognition in the presence of noise, and more particularly to a method for preprocessing speech for use in connection with a speech recognition system.

Speech recognition systems are very sensitive to differences between a training condition which is free of noise and operating conditions in the presence of noise. In particular, speech recognition systems can be trained to recognize specific speech patterns in the absence of noise and are therefore trained on high-quality speech. However, such systems degrade drastically in noisy environments.

Several methods for handling this problem are known, among them are methods of supplementing the acoustic preprocessing of a speech recognizer with a statistical estimator. A statistical estimator as used herein is intended to provide to a speech recognizer

input values or signals which can be assumed to be clean speech information.

The task of designing a statistical estimator for speech recognition is that of defining an optimality criterion that will match the recognizer and of deriving an algorithm to compute the estimator based on this criterion. Defining the optimality criterion is easier for speech recognition than it is for speech enhancement for human listeners, since the signal processing technique is known in the former but not in the latter. For a recognition system which is based on a distance metric, whether for template matching or vector quantization, it is reasonable to assume that the optimality criterion is to minimize the average distortion as measured by that distance metric. Achieving this criterion is frequently computationally infeasible.

With discrete Fourier transform (DFT), filter-bank based systems, the distance measure which is typically used is a weighted Euclidean distance on the cosine transform of the logarithm of the output energy of the filters, often referred to as the "liftered cepstral distance." (The cepstrum in a filter-bank system is defined as a transform of the filter energies.) Achieving this estimation criterion using this distance metric is computationally difficult with additive noise. Published estimation algorithms which have been applied to filter-bank based systems are the minimum mean square error (MMSE) algorithm and the spectral subtraction algorithm, applied to either discrete Fourier transform (DFT) coefficients or filter-bank output energies. (Reference to Porter et al. and Van Compernolle 1 & 2 discussed below.) A basic difference between the multiple-dimensional cepstral distance optimality criterion and the single frequency channel minimum mean square error (MMSE) distance criterion is that the cepstral distance

implies a joint estimation of a feature vector whereas the MMSE distance implies an independent estimation of scalar quantities. Because the speech spectral energies at different frequencies are in fact correlated, use of an independent estimate of individual frequency channels results in suboptimal estimation.

This art presumes a basic familiarity with statistics and Markov processes, as well as familiarity with the state of the art in speech recognition systems using hidden Markov models. By way of example of the state of the art, reference is made to the following patents and publications, which have come to the attention of the inventors in connection with the present invention. Not all of these references may be deemed to be relevant prior art.

Inventor	U.S. Patent No.	Issue Date
Bahl et al.	4,817,156	03/28/89
Levinson et al.	4,587,670	05/06/86
Juang et al.	4,783,804	11/08/88
Bahl et al.	4,741,036	04/26/88

Inventor	Foreign Patent No.	Pub. Date
Sedgwick et al.	EP 240,330	10/07/87

Papers

Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," Proc. IEEE, Vol. 77, No. 2, Feb. 1989.

Nadas et al., " Speech Recognition Using noise-adaptive prototypes," IEEE Trans. on ASSP, Vol. 37, No. 10, Oct. 1989.

Stern et al., "Acoustical pre-processor for robust speech recognition," Proc. DARPA Speech and Natural Language Workshop, Session, October 1989.

Ephraim et al., "Speech Enhancement Using a Minimum Mean-Square Error Short-Time Spectral Estimator," IEEE Trans. ASSP, Vol. 32, pp. 1109-1112 (Dec. 1984).

Ephraim et al., "Speech Enhancement Using a Minimum Mean-Square Error Log-Spectral Amplitude Estimator," IEEE Trans. ASSP, Vol. 33, pp. 443-447 (Apr. 1985).

Porter et al., "Optimal Estimators for Spectral Restoration of Noisy Speech," Proc. ICASSP, Vol. 2, pp. 18A2.1-2.4 (1984).

Van Compernolle, "Noise Adaptation in a Hidden Markov Model Speech Recognition System," Computer Speech and Language, Vol. 3, pp. 151-167, 1989.

Van Compernolle, "Spectral Estimation Using a Log-Distance Error Criterion Applied to Speech Recognition," Proc. ICASSP, Vol. 1, pp. 258-261 (1989).

Gray, "Vector Quantization," The ASSP Magazine, Vol. 1, No. 2, pp. 3-29 (April 1984).

The Rabiner paper is a survey article on Hidden Markov Model applications to speech recognition which is background for understanding the present invention. It does not address the noise problem. It is incorporated herein by reference.

The remaining patents and publications describe other work in the general field.

The Nadas et al. paper describes a method that addresses the noise problem. It is not however a preprocessing method.

The Stern et al. paper describes a method that addresses the noise problem using a preprocessor based on spectral subtraction.

The Van Compernolle (1) is a description of a spectral subtraction type of preprocessor.

The Van Compernolle (2) is a description of a preprocessor using Minimum Mean Square Error independent estimation of single channels.

The Porter et al. and Ephraim et al. papers describe preprocessors for use in speech recognition techniques using the minimum mean square error estimation of various functions of the digital Fourier transform (DFT) coefficients, including the logarithm of the DFT amplitude. These techniques deal only with single DFT coefficients.

Sedgwick et al. describes work at the National Resource Development Corporation on a noise compensation apparatus for speech recognition system, wherein input signals corresponding to levels in frequency spectrum regions are derived and the (Euclidean) distance for noisy input cells is determined to take advantage of noise level during both recognition and training. In both recognition and training processes, signals reaching a microphone are digitized and passed through a filter bank to be separated into frequency channels. In training, a noise estimator and a masker are used with a recognizer to prepare and store probability density functions (PDFs) for each channel partially defining Markov models of words to be recognized. The PDFs are derived only from input signals above noise levels, but derivation is such that the whole of each PDF is represented. In recognition, "distance" measurements on which recognition is based are derived for each channel. If the signal in one channel is above the noise level, then the distance is determined by the recognizer from the negative logarithm of the PDF. If a channel signal is below a noise level, then the distance is determined from the negative logarithm of the cumulative distance of the PDF to the noise level. This publication describes a recognition system with noise compensation, but it does not address preprocessing to address the noise problem.

Bahl et al. '156 describes IBM work on apparatus and method for training the statistics of a

Markov model speech recognizer to a subsequent speaker. Therein, the invention determines label output probabilities at transitions in the Markov models corresponding to the subsequent speaker where there is sparse training data. This patent does not address the noise problem.

Levinson et al. describes AT&T Bell Labs work on a speech recognizer which includes a plurality of stored constrained hidden Markov model reference templates and a set of stored signals representative of prescribed acoustic features of the plurality of reference patterns. This patent does not address the noise problem.

Juang et al. describes further AT&T Bell Labs work on a hidden Markov model speech recognition arrangement. Markov model speech pattern templates are formed by analyzing identified speech patterns to generate frame sequences of acoustic feature signals. This patent does not address the noise problem.

Bahl et al. '036 describes IBM work in a speech recognition system whereby discrimination between similar sounding uttered words is improved by weighting the probability vector data stored for the Markov model representing the reference word sequence of phones. The weighting vector is derived for each reference word by comparing similar sounding utterances using Viterbi alignment and multivariate analysis which maximizes the differences between correct and incorrect recognition multivariate distributions. This patent does not address the noise problem.

None of the prior art teaches how to approximate the cepstral distance optimality criterion in preprocessing of noisy speech in a speech recognition system in a computationally feasible manner.

SUMMARY OF THE INVENTION

In accordance with the invention, a computationally-feasible method is provided for use in preprocessing noisy speech to minimize likelihood of error in estimation for use in a speech recognizer. The computationally-feasible technique, herein called Minimum-Mean-Log-Spectral-Distance (MMLSD) estimation using mixture models and Markov models, comprises the steps of calculating for each vector of speech in the presence of noise corresponding to a single time frame, an estimate of clean speech, where the basic assumptions of the method of the estimator are that the probability distribution of clean speech can be modeled by a mixture of components each representing a different speech class assuming different frequency channels are uncorrelated within each class and that noise at different frequency channels is uncorrelated. (EQ. 11 & FIG. 2). In a further embodiment of the invention, the method comprises the steps of calculating for each sequence of vectors of speech in the presence of noise corresponding to a sequence of time frames, an estimate of clean speech, where the basic assumptions of the method of the estimator are that the probability distribution of clean speech can be modeled by a Markov process assuming different frequency channels are uncorrelated within each state of the Markov process and that noise at different frequency channels is uncorrelated. (EQ. 21 & Fig. 3)

The invention will be better understood upon reference to the following detailed description, taken in conjunction with the accompanying drawings.

BRIEF DESCRIPTION OF THE DRAWINGS

Figure 1 is a block diagram of a speech recognition system incorporating a preprocessor employing a method in accordance with the invention.

Figure 2 is a block diagram of a single frame MMLSD estimator using a mixture model according to a first method in accordance with the invention.

Figure 3 is a flow chart illustrating a computation of a mixture of Gaussians model used in connection with the invention.

Figure 4 is a block diagram of a MMLSD estimator for a sequence of time frames using a Markov model according to a second method in accordance with the invention.

DESCRIPTION OF SPECIFIC EMBODIMENTS

Figure 1 is a block diagram of a speech recognition system 10 incorporating a preprocessing estimator 22 employing a method in accordance with the invention. The system 10 depicted is a filter-bank-based system employing a hidden Markov process recognizer. Alternatively, a template matching system could be employed for speech recognition. The recognition system 10 receives an analog time-domain signal representing speech in the presence of noise at an input 12, which in turn is applied to a feature extractor comprising an analog to digital converter (ADC) 14 and a filter bank 18. The ADC 14 converts the analog signal to digitized speech which are then applied on digital signal lines 16 to a time segmenter 15. The time segmenter 15 segments the digitized signal into time frames for subsequent processing. Output on lines 17 are applied to a filter bank 18. The filter bank 18 categorizes the speech+noise (noisy speech) components into a vector of filter log energies (or more generally spectral log energies) S' with components S'_k for each time frame, each component representing one filter channel of speech information. The vector S' is then applied via lines 20 to a preprocessor 22 which functions as an estimator of the

clean speech. The output of the preprocessor 22 is in the form of a estimate of clean speech, a vector $\hat{S}$.

The vector $\hat{S}$ is optionally applied on lines 24 to an acoustic labeler 26 or directly to a speech recognizer 30.

The preprocessor 22 functions such that all subsequent processing treats input signals as if no noise is present. Three possible speech recognizers can be used with the preprocessor 22 in accordance with the invention. The speech recognizer 30 may be a discrete density Hidden Markov Model (HMM) recognizer with an acoustic labeler using a distance metric for acoustic labeling. Alternatively, it may use a continuous density HMM recognizer which uses a probabilistic algorithm for acoustic labeling but no distance metric. Still further, the speech recognizer 30 may use template matching, such as dynamic time warping, which uses a distance metric for the template matching. The output of the speech recognizer 30 on lines 32 is the recognized speech.

Referring now to Figure 2, there is shown a first embodiment of a preprocessor 22 in accordance with the invention. The preprocessor 22 of Figure 2 computes an estimate, for each time frame, of the vector of clean speech $\hat{S}$ from the vector of noisy speech S' based on:

1) the assumption that the probability distribution of clean speech can be modeled by a mixture of components as a mixture model, each component representing a different speech class assuming different frequency channels are uncorrelated within each class, and based on

2) a conditional probability function of a vector of noisy speech around a vector of clean speech which is based on the assumption that noise at different frequency channels is uncorrelated.

The estimator is a minimum mean square error (MMSE) estimation of the vector S, where the mean square error is in the vector Euclidean sense. The minimum Euclidean distance on the vector S of K filter log-energies yields the following vector estimator, (where boldface indicates a vector):

$$\hat{S} = \int S\, P(S|S')\, dS \qquad (1)$$

Using Bayes' Rule, the estimator is therefore of the form:

$$\hat{S} = \frac{1}{P(S')} \int S\, P(S'|S)\, P(S)\, dS \qquad (2)$$

where the probability for the vector S' is given by:

$$P(S') = \int P(S'|S)\, P(S)\, dS \qquad (3)$$

This estimator is considerably more complex than a single channel-type estimator because it requires integration of K-dimensional probability distributions, e.g., 25 dimensions for 25 frequency channels. In accordance with the invention, approximated models for both the probability of S and the probability of S' around S can be used in the computation, since the assumptions are that the noise is additive and that the vector S is the vector of filter-bank log-energies.

First, the conditional probability of the vector S' around the vector S, or P(S'|S), can be modeled simply as the product of the marginal probabilities, or:

$$P(S'|S) = \prod_{k=1}^{K} P(S'_k|S_k) \qquad (4)$$

since it is assumed that Gaussian noise is uncorrelated in the frequency domain and since the energy value of a given noisy filter S'_k depends only on the clean energy S_k and on the noise level at that frequency (i.e., within the passband of the filter at that frequency). This model is merely an approximation, however, where the passbands of the filters overlap.

The conditional probability $P(S'_k|S_k)$ can be modeled as follows:

Assume that the noise associated with the speech to be estimated can be represented by a stationary, ARMA (autoregressive moving average) stochastic process (i.e, viewed as white noise colored by linear filtering). Therefore, at each time frame, the coefficients of the discrete Fourier transform (DFT) for the noise are uncorrelated complex Gaussian random variables. Assume further that the filter output energy can be approximated by a sum of M coefficients. Finally, assume that the noise spectral power is uniform within the range of summation. With only noise present (i.e., no speech), the sum extends over a total of 2M random variables which are Gaussian, of zero mean and of a variance:

$$\sigma^2 = (N_k)/(2M) \qquad (5)$$

where N_k is the expected value of the noise filter energy. Under these conditions, the variance-normalized filter energy will obey a chi-squared probability distribution (PD) with 2M degrees of freedom.

In the presence of speech and noise, the filter energy is given by:

$$E'_k = \sum_i |DFT_s(i) + DFT_n(i)|^2 \qquad (6)$$

where DFT_s refers to speech coefficients and DFT_n refers to noise coefficients.

The random variable given by Equation 6 divided by Equation 5, or E'_k/σ^2, will then obey a different probability distribution, namely, a probability distribution of noncentral chi-squared with 2M degrees of freedom and noncentral parameter λ.

Thus the conditional probability of the filter energy is given by:

$$P\left[\frac{E'_k}{N_k} \mid \frac{E_k}{N_k}\right] = 2M.\chi^2_{N.C.}\left[\frac{2ME'_k}{N_k},\ 2M, \lambda\right] \qquad (7a)$$

where $$\lambda = \sum_i \frac{|DFT_s(i)|^2}{\sigma^2} = \frac{2ME_k}{N_k} \quad . \qquad (7b)$$

With the following normalized log-energy variable defined,

$$S_k = 10 \log_{10}\left[\frac{E_k}{N_k}\right] ; \quad S'_k = 10 \log_{10}\left[\frac{E'_k}{N_k}\right], \qquad (8)$$

the conditional probability for S'_k is finally given by

$$P(S'_k \mid S_k) = 0.23 \frac{E'_k}{N_k} P\left[\frac{E'_k}{N_k} \mid \frac{E_k}{N_k}\right] \qquad (9)$$

The value S'_k is the filter log-energy value of observed speech in the presence of noise for the k^{th} filter, and value S_k is the filter log-energy value of clean speech.

AUSTRALIAN PATENT OFFICE LS

The conditional probability of S'_k around S_k given by Equation 9 provides the needed "fuzziness" in the probability distribution to account for the uncertainty around the speech element S in the presence of noise.

There may be less than 2M degrees of freedom in actual practice due to deviations from the model assumptions. In particular, the filters are typically defined with trapezoidal frequency windows (rather than box-car or rectangular frequency windows), and the Hamming window of the discrete Fourier transform introduces correlations between the noise DFT coefficients. In addition, the noise spectrum of very broad filters may not be flat within the range of a single filter.

Second, a model is required for the probability distribution (PD) of the clean speech vector P(S). The clean-speech PD cannot be represented in the frequency domain as a product of the marginal probabilities. Rather, this PD can be modeled by a mixture model in the form:

$$P(S) = \sum_{n=1}^{N} C_n P_n(S), \quad P_n(S) = \prod_{k=1}^{K} P_n(S_k) \qquad (10)$$

where C_n are mixture coefficients ~~is a constant~~, and N is the number of mixture components or classes.

This model is based on the idea that the acoustic space can be divided into classes within which the correlation between different frequency channels is significantly smaller than within the space as a whole. The classes can represent either mutually exclusive or overlapping regions of the acoustic space. As hereinafter explained, the estimator for clean speech is given by:

AUSTRALIAN PATENT OFFICE LS

$$\hat{S}_k = \sum_{n=1}^{N} \hat{S}_k|n \cdot P(n|S') , \qquad (11)$$

where index n is a class and where the first term (estimate of S_k around n) is the n^{th} class-conditioned MMSE estimator given in a computationally-feasible form by:

$$\hat{S}_k|n = \frac{1}{P(S'_k|n)} \int S_k \, P(S'_k|S_k) \, P_n(S_k) \, dS_k \qquad (12)$$

where

$$P(S'_k|n) = \int P(S'_k|S_k) \, P_n(S_k) \, dS_k \qquad (13)$$

and the second term (the a posteriori probability of n given the vector S', i.e., that the clean speech vector belonged to the n^{th} class) is given by:

$$P(n|S') = \frac{C_n \, P(S'|n)}{\sum_{n=1}^{N} C_n \, P(S'|n)} \qquad (14)$$

where

$$P(\hat{S}'|n) = \prod_{k=1}^{K} P(\hat{S}'_k|n) . \qquad (15)$$

Thus the estimator can be viewed as a weighted sum of class-conditioned MMSE estimators. Where N=1, the resulting estimator is identical to an MMSE estimator of individual channels.

In order to realize a computationally-feasible estimator according to the invention, it is necessary to employ a mixture model to model P(S) by equation 10. Several implementation of the mixture model may be employed, depending on the application.

AUSTRALIAN PATENT OFFICE

The simplest implementation of the mixture model is by vector quantization. The vector quantization mixture model identifies the classes with rigid-boundary partitions of the acoustic space and creates the partitions by vector quantization. A codebook of size N is created using the Lloyd algorithm. (The Lloyd algorithm is given in the literature in Gray.) The codebook minimizes the distortion as measured by the Euclidean distance, and $P_n(S_k)$ is estimated from a histogram of all speech frames quantized into codeword n.

Another implementation of the mixture model which provides a better fit within computationally-feasible methods is a mixture of Gaussians model. This model is a parametric model for the PDs adjusted so as to maximize the likelihood of observed data given the model.

Parameterization with a mixture of Gaussians model begins with the assumptions that the probabilities $P_n(S_k)$ are Gaussian distributions, each with a mean value of μ_{nk} and a standard deviation σ_{nk}. The maximum likelihood problem thus becomes a problem of estimating Hidden Markov Model (HMM) parameters with continuous densities, where the model comprises a single state with N multivariate Gaussian components of diagonal covariance matrices. Figure 3 illustrates steps for parameter estimation using an iterative procedure as follows:

Make initial estimates of C, μ and σ (Step AA). A reasonable starting point for these "seed" values would be the classification by vector quantization using as values the relative numbers of vectors in each class, their mean and their standard deviation.

Next, perform a loop on all speech frames and for each frame "t" compute the probability $\gamma_n(t)$ according to the equations:

$$\gamma_n(t) = \frac{C_n P_n(S_t)}{\sum_n C_n P_n(S_t)} \qquad (16)$$

where $P_n(\hat{S}_t)$ is computed using the current values of μ and σ (Step AB). The new parameter estimates are then given by the following time averages

$$C_n = <\gamma_n(t)> \qquad (17)$$

$$\mu_{nk} = \gamma_n(t) S_k(t) \qquad (18)$$

$$\sigma^2_{nk} = \gamma_n(t)\{S_k(t) - \mu_{nk}\}^2 \qquad (19)$$

Finally, test for convergence of the total likelihood given by:

$$L = \sum_t \log_{10} [\sum_n C_n P_n(S_t)] \qquad (20)$$

(Step AC). If convergence is not attained, then Step AB is repeated, and the convergence test (Step AC) is repeated until convergence is attained.

The above methods have potential difficulties in practice. First, filter-bank-based systems have filters with passbands that typically overlap, and second, the computational load, even in the foregoing model, may be excessive, particularly for interactive or near real-time recognition systems. With overlapping filters, the assumptions regarding statistical independence of channels are incorrect. While overlapping filters can be replaced with non-overlapping filters, it is also possible to apply a "Broadband" mixture model to reduce the computational load.

AUSTRALIAN PATENT OFFICE LS

With overlapping filters, a K-dimensional vector of filter log-energies can be represented by a vector of less than K dimensions assuming a fewer number of filters with broad yet non-overlapping passbands. Quantization using the VQ mixture model is thereby simplified, and a new, lower-dimensional vector is introduced as a ["] wideband spectral vector [quantization mixture model" as follows]: R, defined as the log energy in a number of frequency bands.

The classification of speech frames proceeds by quantizing the vectors R with a codebook of size N. The mixture components $P_n(S_k)$ of Equation 10 are then estimated based on these classes, and the a posteriori class probability is conditioned on the vector R' rather than on the vector S' as in Equation 11.

$$S_k = \sum_{n=1}^{N} S_k|n \cdot P(n|R') \qquad (21)$$

Equation 21 replaces Equation 11 and the a posteriori class probabilities replacing Equation 14 and 15 are given by:

$$P(n|R') = \frac{C_n P(R'|n)}{\sum_{n=1}^{N} C_n P(R'|n)} \qquad (22)$$

where

$$P(R'|n) = \prod_{j=1}^{J} P\left(R'_j|n\right) \qquad (23)$$

where j is a broadband channel, R_j is the log-energy in channel j, and the total number of bands is J.

$P(R'_j \mid n)$ is computed similarly to Equation (13), where R_j and R'_j replace S_k and S'_k. $P_n(R_j)$ can be either estimated from histograms of the speech frames quantized into codeword n, using vector quantization or it can be modeled by a Gaussian, similar to Gaussian Mixture modeling. $P(R'_j \mid R_j)$ can similarly be modeled.

AUSTRALIAN PATENT OFFICE LS

after $P(S'_k|S_k)$ of Equation 9. The advantage of this method is therefore in the smaller number of integrals to be computed and multiplied in Equation (22), relative to Equation (14).

Referring now to Figure 4, there is illustrated a second embodiment of a preprocessor 22 in accordance with the invention. The preprocessor 22 of Figure 4 computes an estimate, for a sequence of time frames, of the vectors of clean speech, $\hat{S}_0$ to $\hat{S}_T$ from the sequence of vectors of noisy speech, S'_0 to S'_T, based on:

1) the assumption that the probability distribution of clean speech can be modeled by a Markov model, each state of the Markov model representing a different speech class assuming different frequency channels are uncorrelated within each class, and based on

2) a conditional probability function of a vector of noisy speech around a vector of clean speech which is based on the assumption that noise at different frequency channels is uncorrelated.

The estimator is a minimum mean square error (MMSE) estimation of a sequence of vectors S_0 to S_T, where the mean square error is in the vector Euclidean sense. The minimum Euclidean distance estimator of vector S of K filter log-energies given the sequence of noisy vectors S'_0 to S'_T, yields the following vector estimator:

$$\hat{S}_t = \int S_t \, P(S_t|S'_0,\ldots,S'_t\ldots S'_T) \, dS_t. \qquad (24)$$

This vector estimator is computed using the following assumptions. The speech can be modeled by a ~~first order~~ Markov process so that the speech at any time t is in any one of N different states. The output probability distribution (PD) for state "n" is given by:



$$P_n(S)=\prod_{k=1}^{K} P_n(S_k). \qquad (25)$$

The state output probability distributions and the transition probabilities between states are computed by identifying the states with partitions of the acoustic space. These partitions are exactly the same as those used in the vector quantization mixture model or the wideband vector quantization mixture model hereinabove. The transition probabilities are estimated from speech data by counting the number of transitions from one state to another.

With additive noise, the noisy speech is modeled by a Hidden Markov Model (HMM) where the states correspond to the states of the clean speech, and the output probability distributions are given by:

$$P(S'|n) = \prod_{k=1}^{K} P(S'_k|n) . \qquad (26)$$

where $P(S'_k|n)$ is given by

$$P(S'_k|n) = \int P(S'_k|S_k)\ P_n(S_k)\ dS_k \qquad (27)$$

and where $P_n(S_k)$ is the output probability distribution of the k^{th} component of the clean speech vector S for state "n". The probability of S'_k given S_k is computed using equations 7-9 hereinabove.

Given all of these assumptions, the estimator of Figure 4 is given by:

$$S_k = \sum_n \hat{S}_k|n \cdot P(n|S'_0,\ldots,S'_t\ldots S'_T). \qquad (28)$$

The probability of state "n" given the sequence of noisy speech vectors in Equation 28 is computed by the forward-backward algorithm applied to the HMM of



the noisy speech defined above. (The forward-backward algorithm is for example given in the Rabiner (1989) review article, referenced above, which is incorporated herein by reference and made a part hereof.)

~~A sample of one implementation of the six~~ different preprocessor estimators in ~~accordance~~ with the present invention is ~~presented~~ in Appendix A attached ~~hereto. The~~ appendix is source code written ~~in the LISP programming language.~~

The invention has now been described with reference to specific embodiments. Other embodiments will be apparent to those of ordinary skill in the art. It is therefore not intended that this invention be limited, except as indicated in the appended claims.

AUSTRALIAN PATENT OFFICE

THE CLAIMS DEFINING THE INVENTION ARE AS FOLLOWS:

1. In a speech recognition system for responding to signals representative of digitized speech, a method for preprocessing a noisy speech signal to estimate a clean speech signal, comprising the steps of:

categorizing each time-invariant segment of digitized speech and noise into a noisy speech vector, each element of said noisy speech vector representing one frequency channel of speech information in the presence of noise, said noise being assumed to be additive, time invariant and uncorrelated to the frequency domain;

estimating a clean speech signal comprising a clean speech vector for each said element from said noisy speech vector by

(1) determining the probability distribution of clean speech by applying the noisy speech vector to a mixture model composed of a mixture of components, each component representing a different speech class assuming different frequency channels are uncorrelated within each class,

(2) modelling the conditional probability function of a vector of noisy speech around a vector of clean speech as uncorrelated at different frequency channels; and

storing the estimated clean speech signal in memory.

2. The method according to claim 1 wherein said estimating step comprises:

estimating filter log spectral energies on a multiple-frequency channel probability according to the function:

$$S_k = \sum_{n=1}^{N} S_k|n \cdot P(n|S')$$

where index n is a class and where the estimate of S_k around n is the nth class-conditioned MMSE estimator given by:

$$S_k|n = \frac{1}{P(S'_k|n)} \int S_k P(S'_k|S_k) P_n(S_k) dS_k$$

where

$$P(S'_k|n) = \int P(S'_k|S_k) P_n(S_k) dS_k .$$

3. The method according to claim 2 wherein said estimating step comprises applying a minimum mean-log-spectral-distance estimation to said vector of noisy speech to obtain an estimation of Euclidean distance.

4. The method according to claim 1 wherein said conditional probability of a noisy speech vector around a clean speech vector P(S'k Sk) is modeled as follows:

1) noise associated with said speech to be estimated is assumed to be a stationary, autoregressive moving average stochastic process such that, at each time frame, coefficients of a discrete Fourier transform (DFT) for said noise are uncorrelated complex Gaussian random variables;

2) filter output energy is assumed to be approximated by a sum of M coefficients;

3) noise spectral power is assumed to be uniform within a range of passbands of filters;

such that, in the presence of noise alone, said summation extends over a total of 2M random variables which are Gaussian, of zero mean and of a variance:

$$\sigma^2 = (N_k)/(2M)$$

where N_k is the expected value of the noise filter energy wherein variance-normalized filter energy obeys a chi-squared probability distribution (PD) with 2M degrees of freedom, and

such that, in the presence of speech and noise, filter energy is given by:

$$E'_k = \sum_i |DFT_s(i) + DFT_n(i)|^2$$



where DFT_x refers to speech coefficients and DFT_n refers to noise coefficients, wherein said variance-normalized filter obeys a probability distribution of noncentral chi-squared with 2M degrees of freedom and noncentral parameter λ;

in order that the conditional probability of the filter energy is given by:

$$P\left[\frac{E'_k}{N_k}\middle|\frac{E_k}{N_k}\right] = 2M \cdot \chi^2_{N.C.}\left[\frac{2ME'_k}{N_k}, 2M, \lambda\right]$$

where

$$\lambda = \sum_i \frac{\left|DFT_s(i)\right|^2}{\sigma^2} = \left[\frac{2ME_k}{N_k}\right]$$

so that, with the following normalized log-energy variable defined as:

$$S_k = 10\log_{10}\left[\frac{E_k}{N_k}\right];\ S'_k = 10\log_{10}\left[\frac{E'_k}{N_k}\right]$$

the conditional probability for S'_k is given by

$$P = \left(S'_k \middle| S_k\right) = 0.23\frac{E'_k}{N_k} P\left[\frac{E'_k}{N_k}\middle|\frac{E_k}{N_k}\right]$$

5. The method according to claim 1 wherein said mixture model is implemented as a vector quantization mixture model where classes are identified with rigid boundary partitions of an acoustic space and partitions are created by vector quantization.

6. The method according to claim 1 wherein said mixture model is implemented as a mixture of Gaussians mixture model with the assumptions that the probabilities Pn(Sk) are Gaussian distributions and the mean and standard deviations

of these Gaussians are adjusted so as to maximize the likelihood of speech data given the mixture model.

AUSTRALIAN PATENT OFFICE LS

7. The method according to claim 2 wherein the vector S' of said mixture model is replaced by a vector R', where the vector R' is a lower dimensional vector of log spectral energies in wide frequency bands and wherein the probability of class n conditioned on S' is replaced by the probability of new class n conditioned on R' such that new class n components refer to classes in a mixture model for a vector R.

8. The method according to claim 7 wherein the mixture model for vector R is implemented as a vector quantization mixture model where classes are identified with rigid boundary partitions of acoustic space and the partitions are created by vector quantization and wherein S k in the expression for the estimator for clean speech:

$$S_k = \sum_{n=1}^{N} S_k|n \cdot P(n|R')$$

is conditioned on new class n.

9. The method according to claim 7 wherein the mixture model for vector R is implemented as a mixture of Gaussians mixture model with the assumptions that the probabilities $P_n(R_j)$ are Gaussian distributions and the mean and standard deviations of these Gaussians are adjusted so as to maximize the likelihood of speech data given the mixture model and wherein S_k in the expression for the estimator for clean speech:

$$S_k = \sum_{n=1}^{N} S_k|n \cdot P(n|R')$$

is conditioned on new class n.

10. In a speech recognition system for responding to signals representative of digitized speech, a method for preprocessing speech in the presence of noise to produce an estimated clean speech signal, comprising the steps of:

providing a data base of clean speech;

training a Markov model with said clean speech data base, each state of said Markov model representing a different speech class assuming different frequency channels are uncorrelated within each class, to produce parameters for said Markov model;

categorizing each time-invariant segment of digitized speech and noise into a noisy speech vector, each element of said noisy speech vector representing one frequency channel of speech information in the presence of noise, said noise being assumed to be frequency independent and time invariant;

estimating a clean speech signal comprising a clean speech vector for each vector of a sequence of said noisy speech vectors S'_O to S'_T by

1) determining a probability distribution of the clean speech vector by applying the noisy speech vector to the Markov model using said parameters, and

2) modelling a conditional probability function of a vector of noisy speech around a vector of clean speech as uncorrelated at different frequency channels.

11. The method according to claim 10 wherein said estimating step comprises:

estimating filter log spectral energies on a multiple-frequency channel probability according to

$$S_k = \sum_n S_k|n \cdot P(n|S'_O \ldots, S'_t \ldots S'_T)$$

where index n is a state and where the estimate of S_k around n is the n^{th} MMSE estimator given by:

$$S_k|n = \frac{1}{P(S'_k|n)} \int S_k P(S'_k|S_k) P_n(S_k) dS_k$$

where

$$P(S'_k|n) = \int P(S'_k|S_k) P_n(S_k) dS_k$$

where a Hidden Markov model is assumed for the speech, and the Hidden Markov state output probabilities are specified by:

$$P(S'|n) = \prod_{k=1}^{K} P(S'_k|n)$$

and where a forward-backward algorithm is applied to compute $P(n|S'_0 \ldots, S'_t \ldots S'_T)$.

12. The method according to claim 10 wherein said conditional probability of a noisy speech vector around a clean speech vector $P(S'_k\ S_k)$ is modeled as follows:

1) noise associated with said speech to be estimated is assumed to be a stationary, autoregressive moving average stochastic process such that, at each time frame, coefficients of a discrete Fourier transform (DFT) for said noise are uncorrelated complex Gaussian random variables;

2) filter output energy is assumed to be approximated by a sum of M coefficients;

3) noise spectral power is assumed to be uniform within a range of summation;

such that, in the absence of noise, said summation extends over a total of 2M random variables which are Gaussian, of zero means and of a variance:

$$\sigma^2 = (N_k)/(2M)$$

where N_k is the expected value of the noise filter energy wherein variance-normalized filter energy obeys a chi-squared probability distribution (PD) with 2M degrees of freedom,

and such that, in the presence of speech and noise, filter energy is given by:

$$E'_k = \sum_i |DFT_s(i) + DFT_n(i)|^2$$

where DFT_s refers to speech coefficients and DFT_n refers to noise coefficients, wherein said variance-normalized filter obeys a probability

distribution of noncentral chi-squared with 2M degrees of freedom and noncentral parameter λ p

in order that the conditional probability of the filter energy is given by:

$$P\left[\frac{E'_k}{N_k}\middle|\frac{E_k}{N_k}\right]=2M\cdot\chi^2_{N.C.}\left[\frac{2ME'_k}{N_k},2M,\lambda\right]$$

where

$$\lambda=\sum_i\frac{\left|DFT_s(i)\right|^2}{\sigma^2}=\left[\frac{2ME_k}{N_k}\right]$$

so that, with the following normalized log-energy variable defined as:

$$S_k=10\log_{10}\left[\frac{E_k}{N_k}\right];\ S'_k=10\log_{10}\left[\frac{E'_k}{N_k}\right]$$

the conditional probability for S'_k is given by

$$P=\left(S'_k\middle|S_k\right)=0.23\frac{E'_k}{N_k}P\left[\frac{E'_k}{N_k}\middle|\frac{E_k}{N_k}\right]$$

13. The method according to claim 10 wherein said Markov model is implemented as a vector quantization Markov model where states are identified with rigid boundary partitions of an acoustic space and partitions are created by vector quantization.

14. The method according to claim 11 wherein the vector S' of said mixture model is replaced by a vector R', where the vector R' is a lower dimensional vector of log spectral energies in wide frequency bands and wherein the probability of state n conditioned on sequence of S'(t) where t is from zero to T is replaced by the probability of new state n conditioned on sequence of R'(t) where t is from zero to T such that new state n components refer to states in a Markov model for a vector R.

AUSTRALIAN PATENT OFFICE

15. The method according to claim 14 wherein the mixture model for vector R is implemented as a vector quantization Markov model where states are identified with rigid boundary partitions of an acoustic space and partitions are created by vector quantization and S_k is conditioned on new state n.

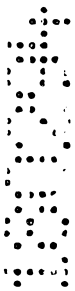
16. The method according to claim 1 and 10, substantially as herein described with reference to the acompanying drawings.

DATED this 17th day of February, 1994.

SRI INTERNATIONAL

WATERMARK PATENT & TRADEMARK ATTORNEYS
THE ATRIUM
290 BURWOOD ROAD
HAWTHORN VICTORIA 3122
AUSTRALIA

SKP:RCS:KP/JL VAX doc 047 AU7487591.WPC





74875|91

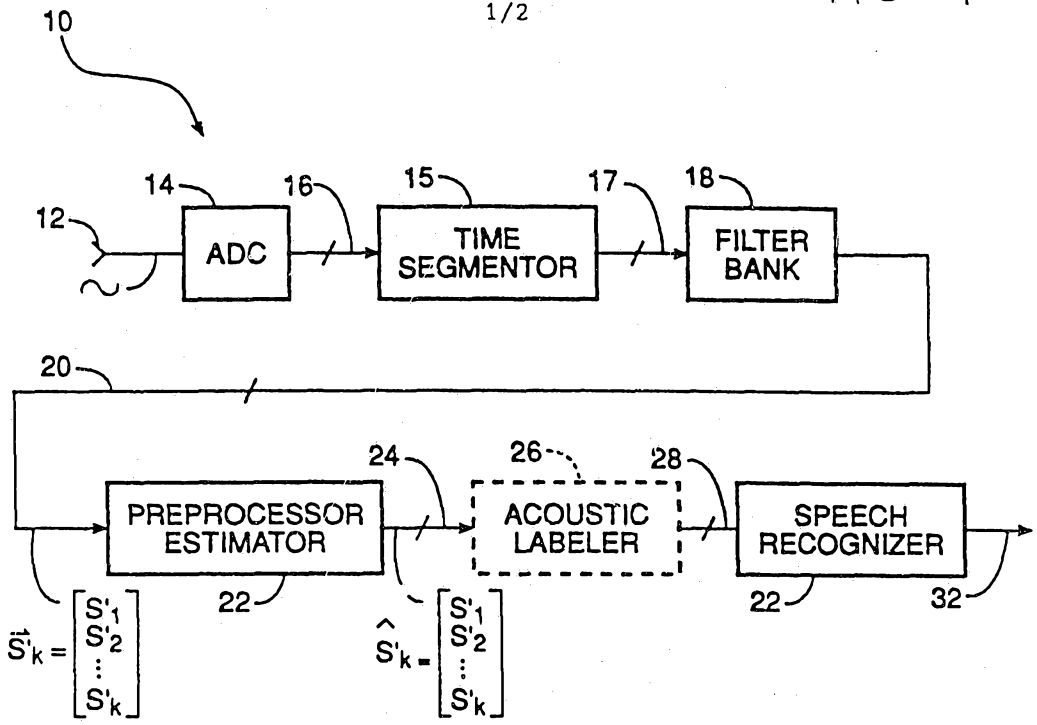


FIG. 1

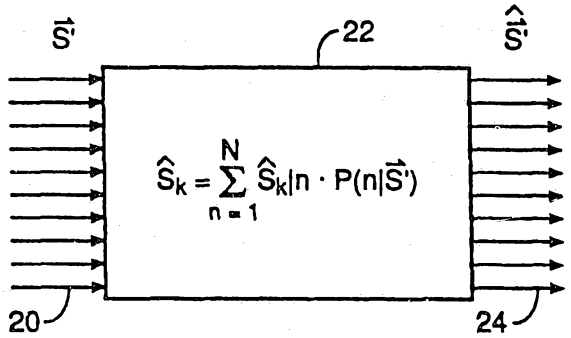


FIG. 2

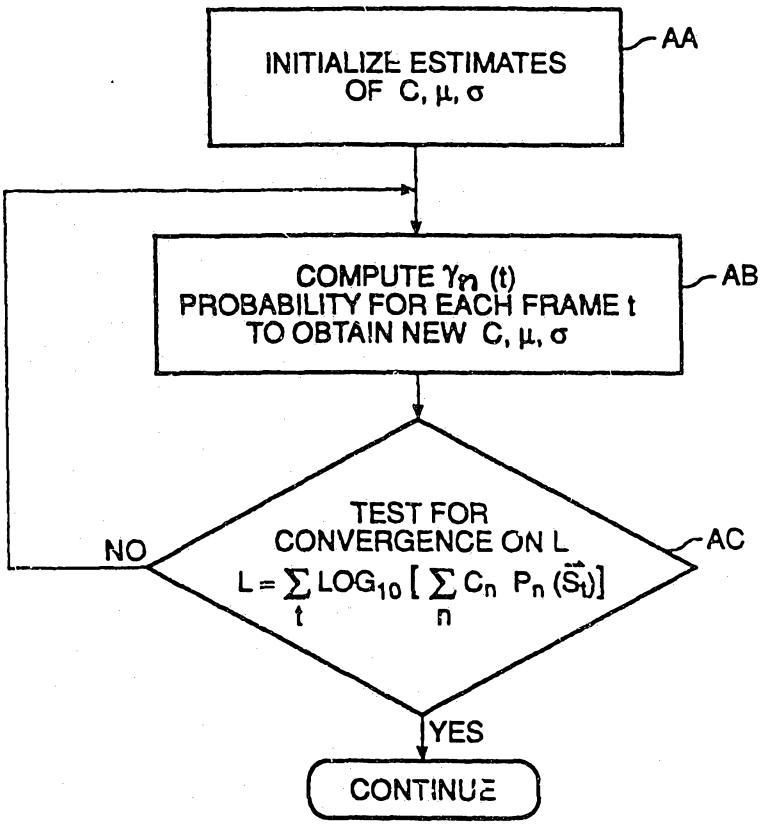

INITIALIZE ESTIMATES OF C, μ, σ
AA
COMPUTE γ_n (t) PROBABILITY FOR EACH FRAME t TO OBTAIN NEW C, μ, σ
AB
TEST FOR CONVERGENCE ON L
$L = \sum_t \text{LOG}_{10} [\sum_n C_n P_n (\vec{S}_t)]$
AC
NO
YES
CONTINUE

FIG. 3

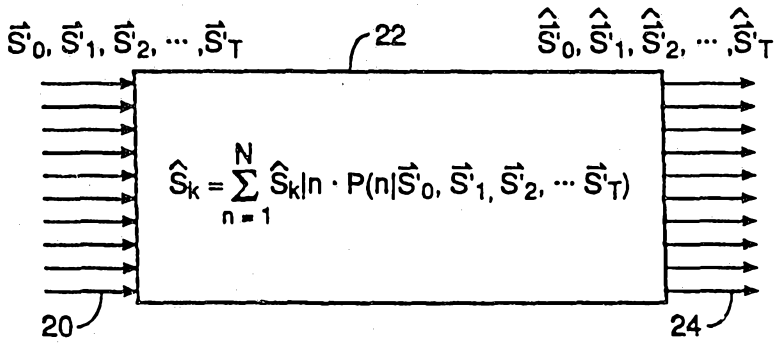

$\vec{S}'_0, \vec{S}'_1, \vec{S}'_2, \cdots, \vec{S}'_T$
22
$\hat{\vec{S}}'_0, \hat{\vec{S}}'_1, \hat{\vec{S}}'_2, \cdots, \hat{\vec{S}}'_T$
$\hat{S}_k = \sum_{n=1}^{N} \hat{S}_k|n \cdot P(n|\vec{S}'_0, \vec{S}'_1, \vec{S}'_2, \cdots \vec{S}'_T)$
20
24

FIG. 4

INTERNATIONAL SEARCH REPORT

International Application No PCT/US 91/01333

I. CLASSIFICATION OF SUBJECT MATTER (if several classification symbols apply, indicate all) [6]

According to International Patent Classification (IPC) or to both National Classification and IPC

IPC[5]: G 10 L 3/00

II. FIELDS SEARCHED

Minimum Documentation Searched [7]

Classification System	Classification Symbols
IPC[5]	G 10 L

Documentation Searched other than Minimum Documentation to the Extent that such Documents are Included in the Fields Searched [8]

III. DOCUMENTS CONSIDERED TO BE RELEVANT [9]

Category [10]	Citation of Document, [11] with indication, where appropriate, of the relevant passages [12]	Relevant to Claim No. [13]
A	ICASSP 88, International Conference on Acoustics, Speech, and Signal Processing, New York, 11-14 April 1988, vol. 1, IEEE, Y. Ephraim et al.: "On the application of hidden Markov models for enhancing noisy speech", pages 533-536 see the whole article --	1-6,10-13
A	IEEE Transactions on Acoustics, Speech and Signal Processing, vol. ASSP-33, no. 2, April 1985, IEEE, (New York, US), Y. Ephraim et al.: "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator", pages 443-445 see page 443, section II, "Derivation of MMSE log-STSA estimator" cited in the application --	2,3,11

./.

* Special categories of cited documents: [10]

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

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

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

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

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

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

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

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

"&" document member of the same patent family

IV. CERTIFICATION

Date of the Actual Completion of the International Search	Date of Mailing of this International Search Report
17th June 1991	25 JUL 1991
International Searching Authority	**Signature of Authorized Officer**
EUROPEAN PATENT OFFICE	WISST TAZELAAR

III. DOCUMENTS CONSIDERED TO BE RELEVANT (CONTINUED FROM THE SECOND SHEET)

Category *	Citation of Document, " with indication, where appropriate, of the relevant passages	Relevant to Claim No.
A	IEEE Transactions on Acoustics, Speech and Signal Processing, vol. ASSP-32, no. 6, December 1984, IEEE, (New York, US), Y. Ephraim et al.: "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator", pages 1109-1121 see page 1117, "Maximum likelihood estimation approach" cited in the application	6
A	IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 37, no. 10, October 1989, IEEE, (New York, US), A. Nádas et al.: "Speech recognition using noise-adaptive prototypes", pages 1495-1502 see page 1495, abstract and introduction	1,10
A	IEEE Transactions on Information Theory, vol. 34, no. 4, July 1988, IEEE, Y. Ephraim et al.: "A unified approach for encoding clean and noisy sources by means of waveform and autoregressive model vector quantization", pages 826-834 see page 831, VI "Examples"	5,6,13
A	EP, A, 0240330 (NATIONAL RESEARCH) 7 October 1987 cited in the application	
A	Computer Speech and Language, vol. 3, April 1989, Academic Press Ltd, D. Van Compernolle: "Noise adaptation in a hidden Markov model speech recognition system", pages 151-167 cited in the application	

ANNEX TO THE INTERNATIONAL SEARCH REPORT ON INTERNATIONAL PATENT APPLICATION NO. US 9101333 SA 45426

This annex lists the patent family members relating to the patent documents cited in the above-mentioned international search report. The members are as contained in the European Patent Office EDP file on 16/07/91
The European Patent Office is in no way liable for these particulars which are merely given for the purpose of information.

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
EP-A- 0240330	07-10-87	GB-A,B 2188764	07-10-87
		GB-A- 2212962	02-08-89
		JP-A- 62265700	18-11-87
		US-A- 4905286	27-02-90