Description

Technical Field

This invention relates to classifying samples representing a real time process into groups with each group corresponding to a state of the real time process. In particular, the classifying is done in real time as each sample is generated using statistical techniques.

Background and Problem

In many real time processes, a problem exists in attempting to estimate the present state of the process in a changing environment from present and past samples of the process. One example of such a process is the generation of speech by the human vocal tract. The sound produced by the vocal tract can have a fundamental frequency - voiced state or no fundamental frequency - unvoiced state. Further, a third state may exist if no sound is being produced - silence state. The problem of determining these three states is referred to as the voicing/silence decision. In low bit rate voice coders, degradation of voice quality is often due to inaccurate voicing decisions. The difficulty in correctly making these voicing decisions lies in the fact that no single speech parameter or classifier can reliably distinguish voiced speech from unvoiced speech. In order to make the voicing decision, it is known in the art to combine multiple speech classifiers in the form of a weighted sum. Such a method is illustrated in D. P. Prezas, et al., "Fast and Accurate Pitch Detection Using Pattern Recognition and Adaptive Time-Domain Analysis," Proc. IEEE Int. Conf. Acoust., Speech and Signal Proc., Vol. 1, pp. 109-112, April 1986. As described in that article, a frame of speech is declared voiced if a weighted sum of speech classifiers is greater than a specified threshold; and unvoiced otherwise. Mathematically, this relationship may be expressed as $a'x + b > 0$ where "a" is a vector comprising the weights, "x" is a vector comprising the classifiers, and "b" is a scalar representing the threshold value. The weights are chosen to maximize performance on a training set of speech where the voicing of each frame is known. These weights form a decision rule which provides significant speech quality improvements in speech coders compared to those using a single parameter.

A problem associated with the fixed weighted sum method is that it does not perform well when the speech environment changes. Such changes in the speech environment may be a result of a telephone conversation being carried on in a car via a mobile telephone or maybe due to different telephone transmitters. The reason that the fixed weighted sum methods do not perform well in changing environments is that many speech classifiers are influenced by background noise, non-linear distortion, and filtering. If voicing is to be determined for speech with characteristics different from that of the training set, the weights, in general, will not yield satisfactory results.

One method for adapting the fixed weighted sum method to changing speech environment is disclosed in the paper of J.P. Campbell et al., "Voice/Unvoiced Classification of Speech with Application to the U.S. Government LPC-10E Algorithm," IEEE International Conference on Acoustics, Speech and Signal Processing, 1986, Tokyo, Vol. 9.11.4, pp. 473-476. This paper discloses the utilization of different sets of weights and threshold values each of which has been predetermined from the same set of training data with different levels of white noise being added to the training data for each set of weights and threshold value. For each frame, the speech samples are processed by a set of weights and a threshold value after the results of one of these sets is chosen on the basis of the value of a signal-to-noise-ratio, SNR. The range of possible values that the SNR can have is subdivided into subranges with each subrange being assigned to one of the sets. For each frame, the SNR is calculated; the subrange is determined; and then, the detector associated with this subrange is used to determine whether the frame is unvoiced/voiced. The problem with this method is that it is only valid for the training data plus white noise and cannot adapt to a wide range of speech environments and speakers. Therefore, there exists a need for a voiced detector that can reliably determine whether speech is unvoiced or voiced for a varying environment and different speakers.

B.S. Atal and L.R. Rabiner, in "A Pattern Recognition Approach to Voiced-Unvoiced-Silence Classification with Applications to Speech Recognition", IEEE Trans. Acoust., Speech Signal Processing, vol. ASSP-24, no. 3, pages 201-212, June 1976, disclose a process for distinguishing between silence, voiced speech and unvoiced speech from a number of classifiers using a minimum-distance criterion in which the distance is defined by means of mean values and covariances of the classifiers for the three classes, determined from a manually classified set of training utterances.

P. de Souza, in "A Statistical Approach to the Design of an Adaptive Self-Normalizing Silence Detector", IEEE Trans. Acoust. Speech, Signal Processing, vol. ASSP-31, no. 3, pages 678-684, June 1983, discloses a process for detecting silence in which the first half-second of an input signal is assumed to be

silent and means and covariances of the classifiers are calculated from this small sample. These are then used to detect, using a statistical test, a subsequent half-second of silence, which is then added to the original sample and the means and covariances updated, and so on until ten half-seconds of silence have been collected. The detector then returns to the start of the signal, using the means and covariances derived from the ten half-seconds of silence.

Solution.

The present invention accordingly provides an apparatus analysing silence free speech for making a voice decision in speech signals as claimed in claim 1.

The above described problem is solved and a technical advance is achieved by an apparatus that is responsive to real time samples from a physical process to determine statistical distributions for plurality of process states and from those distributions to establish decision regions. The latter regions are used to determine the present process state as each process sample is generate. For use in making a voicing decision, the apparatus adapts to a changing speech environment by utilizing the statistics of classifiers of the speech. Statistics are based on the classifiers and are used to modify the decision regions used in the voicing decision. Advantageously, the apparatus estimates statistical distributions for both voiced and unvoiced frames and uses those statistical distributions for determining decision regions. The latter regions are then used to determine whether a present speech frame is voiced or unvoiced.

Advantageously, a voiced detector calculates the probability that the present speech frame is unvoiced, the probability that the present speech frame is voiced, and an overall probability that any frame will be unvoiced. Using these three probabilities, the detector then calculates the probability distribution of unvoiced frames and the probability distribution of voiced frames. In addition, the calculation for determining the probability that the present speech frame is voiced or unvoiced is performed by doing a maximum likelihood statistical operation. Also, the maximum likelihood statistical operation is responsive to a weight vector and a threshold value in addition to the probabilities. In another embodiment, the weight vector and threshold value are adaptively calculated for each frame. This adaptive calculation of the weight vector and the threshold value allows the detector to rapidly adapt to changing speech environments.

Advantageously, an apparatus for determining the presence of the fundamental frequency in frames of speech has a circuit responsive to a set of classifiers representing the speech attributes of a speech frame for calculating a set of statistical parameters. A second circuit is responsive to the calculated set of parameters defining the statistical distributions to calculate a set of weights each associated with one of the classifiers. Finally, a third circuit in response to the calculated set of weights and classifiers and the set of parameters determines the presence of the fundamental frequency in the speech frame or as it is commonly expressed makes the unvoiced/voiced decision.

Advantageously, the second circuit also calculates a threshold value and a new weight vector and communicates these values to the first circuit that is responsive to these values and a new set of classifiers for determining another set of statistical parameters. This other set of statistical parameters is then used to determine the presence of the fundamental frequency for the next frame of speech.

Advantageously, the first circuit is responsive to the next set of classifiers and the new weight vector and threshold value to calculate the probability that the next frame is unvoiced, the probability that the next frame is voiced, and the overall probability that any frame will be unvoiced. These probabilities are then utilized with a set of values giving the average of classifiers for past and present frames to determine the other set of statistical parameters.

The method for determining a voicing decision is performed by the following steps: estimating statistical distributions for voiced and unvoiced frames, determining decision regions representing voiced and unvoiced speech in response to the statistical distributions, and making the voicing decision in response to the decision regions and a present speech frame. In addition, the statistical distributions are calculated from the probability that the present speech frame is unvoiced, the probability that the present speech frame is voiced, and the overall probability that any frame will be unvoiced. These three probabilities are calculated as three sub-steps of the step of determining the statistical distributions.

Brief Description of the Drawing

The invention may be better understood from the following detailed description which when read with the reference to the drawing in which:

FIG. 1 is a block diagram of an apparatus using the present invention;

FIG. 2 illustrates, in block diagram form, the present invention;

FIGS. 3 and 4 illustrate, in greater detail, the functions performed by statistical voiced detector 103 of FIG. 2; and

FIG. 5 illustrates, in greater detail, functions performed by block 340 of FIG. 4.

Detailed Description

FIG. 1 illustrates an apparatus for performing the unvoiced/voiced decision operation using as one of the voiced detectors a statistical voiced detector which is the subject of this invention. The apparatus of FIG. 1 utilizes two types of detectors: discriminant and statistical voiced detectors. Statistical voiced detector 103 is an adaptive detector that detects changes in the voice environment and modifies the weights used to process classifiers coming from classifier generator 101 so as to more accurately make the unvoiced/voiced decision. Discriminant voice detector 102 is utilized during initial start up or rapidly changing voice environment conditions when statistical voice detector 103 has not yet fully adapted to the initial or new voice environment.

Consider now the overall operation of the apparatus illustrated in FIG. 1. Classifier generator 101 is responsive to each frame of speech to generate classifiers which advantageously may be the log of the speech energy, the log of the LPC gain, the log area ratio of the first reflection coefficient, and the squared correlation coefficient of two speech segments one frame long which are offset by one pitch period. The calculation of these classifiers involves digitally sampling analog speech, forming frames of the digital samples, and processing those frames and is well known in the art. Generator 101 transmits the classifiers to detectors 102 and 103 via path 106.

Detectors 102 and 103 are responsive to the classifiers received via path 106 to make unvoiced/voiced decisions and transmit these decisions via paths 107 and 110, respectively, to multiplexer 105. In addition, the detectors determine a distance measure between voiced and unvoiced frames and transmit these distances via paths 108 and 109 to comparator 104. Advantageously, these distances may be Mahalanobis distances or other generalized distances. Comparator 104 is responsive to the distances received via paths 108 and 109 to control multiplexer 105 so that the latter multiplexer selects the output of the detector that is generating the largest distance.

FIG. 2 illustrates, in greater detail, statistical voiced detector 103. For each frame of speech, a set of classifiers also referred to as a vector of classifiers is received via path 106 from classifier generator 101. Silence detector 201 is responsive to these classifiers to determine whether or not speech is present in the present frame. If speech is present, detector 201 transmits a signal via path 210. If no speech (silence) is present in the frame, then only subtractor 207 and U/V determinator 205 are operational for that particular frame. Whether speech is present or not, the unvoiced/voiced decision is made for every frame by determinator 205.

In response to the signal from detector 201, classifier averager 202 maintains an average of the individual classifiers received via path 106 by averaging in the classifiers for the present frame with the classifiers for previous frames. If speech (non-silence) is present in the frame, silence detector 201 signals statistical calculator 203, generator 206, and averager 202 via path 210.

Statistical calculator 203 calculates statistical distributions for voiced and unvoiced frames. In particular, calculator 203 is responsive to the signal received via path 210 to calculate the overall probability that any frame is unvoiced and the probability that any frame is voiced. In addition, statistical calculator 203 calculates the statistical value that each classifier would have if the frame was unvoiced and the statistical value that each classifier would have if the frame was voiced. Further, calculator 203 calculates the covariance matrix of the classifiers. Advantageously, that statistical value may be the mean. The calculations performed by calculator 203 are not only based on the present frame but on previous frames as well. Statistical calculator 203 performs these calculations not only on the basis of the classifiers received for the present frame via path 106 and the average of the classifiers received path 211 but also on the basis of the weight for each classifiers and a threshold value defining whether a frame is unvoiced or voiced received via path 213 from weights calculator 204.

Weights calculator 204 is responsive to the probabilities, covariance matrix, and statistical values of the classifiers for the present frame as generated by calculator 203 and received via path 212 to recalculate the values used as weight vector a, for each of the classifiers and the threshold value b, for the present frame. Then, these new values of a and b are transmitted back to statistical calculator 203 via path 213.

Also, weights calculator 204 transmits the weights and the statistical values for the classifiers in both the unvoiced and voiced regions via path 214, determinator 205, and path 208 to generator 206. The latter generator is responsive to this information to calculate the distance measure which is subsequently transmitted via path 109 to comparator 104 as illustrated in FIG. 1.

U/V determinator 205 is responsive to the information transmitted via paths 214 and 215 to determine whether or not the frame is unvoiced or voiced and to transmit this decision via path 110 to multiplexer 105 of FIG. 1.

Consider now in greater detail the operation of each block illustrated in FIG. 2 which is now given in terms of vector and matrix mathematics. Averager 202, statistical calculator 203, and weights calculator 204 implement an improved EM algorithm similar to that suggested in the article by N. E. Day entitled "Estimating the Components of a Mixture of Normal Distributions", Biometrika, Vol. 56, no. 3, pp. 463-474, 1969. Utilizing the concept of a decaying average, classifier averager 202 calculates the average for the classifiers for the present and previous frames by calculating following equations 1, 2, and 3:

$$n = n+1 \text{ if } n < 2000 \quad (1)$$

$$z = 1/n \quad (2)$$

$$X_n = (1-z)\, X_{n-1} + z x_n \quad (3)$$

x_n is a vector representing the classifiers for the present frame, and n is the number of frames that have been processed up to 2000. z represents the decaying average coefficient, and X_n represents the average of the classifiers over the present and past frames. Statistical calculator 203 is responsive to receipt of the z, x_n and X_n information to calculate the covariance matrix, T, by first calculating the matrix of sums of squares and products, Q_n, as follows:

$$Q_n = (1-z)\, Q_{n-1} + z\, x_n\, x'_n . \quad (4)$$

After Q_n has been calculated, T is calculated as follows:

$$T = Q_n - X_n\, X'_n . \quad (5)$$

The means are subtracted from the classifiers as follows:

$$x_n = x_n - X_n \quad (6)$$

Next, calculator 203 determines the probability that the frame represented by the present vector x_n is unvoiced by solving equation 7 shown below where, advantageously, the components of vector a are initialized as follows: component corresponding to log of the speech energy equals 0.3918606, component corresponding to log of the LPC gain equals -0.0520902, component corresponding to log area ratio of the first reflection coefficient equals 0.5637082, and component corresponding to squared correlation coefficient equals 1.361249; and b initially equals -8.36454:

$$P(u \mid x_n) = \frac{1}{1 + \exp(a' x_n + b)} . \quad (7)$$

After solving equation 7, calculator 203 determines the probability that the classifiers represent a voiced frame by solving the following:

$$P(v|x_n) = 1 - P(u|x_n) \quad (8)$$

Next, calculator 203 determines the overall probability that any frame will be unvoiced by solving equation 9 for p_n:

$$p_n = (1-Z)\, p_{n-1} + z\, P(u|x_n) . \quad (9)$$

After determining the probability that a frame will be unvoiced, calculator 203 then determines two vectors, u and v, which give the mean values of each classifier for both unvoiced and voiced type frames. Vectors u and v are the statistical averages for unvoiced and voiced frames, respectively. Vector u, statistical average unvoiced vector, contains the mean values of each classifier if a frame is unvoiced; and

vector v, statistical average voiced vector, gives the mean value for each classifier if a frame is voiced. Vector u for the present frame is solved by calculating equation 10, and vector v is determined for the present frame by calculating equation 11 as follows:

$$u_n = (1-z)\, u_{n-1} + z\, x_n\, P(u|x_n)/p_n - z x_n \qquad (10)$$

$$v_n = (1-z)\, v_{n-1} + z\, x_n\, P(v|x_n)/(1-p_n) - z x_n \qquad (11)$$

Calculator 203 now communicates the u and v vectors, T matrix, and probability p to weights calculator 204 via path 212.

Weights calculator 204 is responsive to this information to calculate new values for vector a and scalar b. These new values are then transmitted back to statistical calculator 203 via path 213. This allows detector 103 to adapt rapidly to changing environments. Advantageously, if the new values for vector a and scalar b are not transmitted back to statistical calculator 203, detector 103 will continue to adapt to changing environments since vectors u and v are being updated. As will be seen, determinator 205 uses vectors u and v as well as vector a and scalar b to make the voicing decision. If n is greater than advantageously 99, vector a and scalar b are calculated as follows. Vector a is determined by solving the following equation:

$$a = \frac{T^{-1}(v_n - u_n)}{1 - p_n(1-p_n)\,(u_n - v_n)'\, T^{-1}(u_n - v_n)} . \qquad (12)$$

Scalar b is determined by solving the following equation:

$$b = -\tfrac{1}{2}\, a'(u_n + v_n) + \log[(1-p_n)/p_n] . \qquad (13)$$

After calculating equations 12 and 13, weights calculator 204 transmits vectors a, u, and v to block 205 via path 214. If the frame contained silence only equation 6 is calculated.

Determinator 205 is responsive to this transmitted information to decide whether the present frame is voiced or unvoiced. If the element of vector $(v_n - u_n)$ corresponding to power is positive, then, a frame is declared voiced if the following equation is true:

$$a'x_n - a'(u_n + v_n)/2 > 0 ; \qquad (14)$$

or if the element of vector $(v_n - u_n)$ corresponding to power is negative, then, a frame is declared voiced if the following equation is true:

$$a'x_n - a'(u_n + v_n)/2 < 0 . \qquad (15)$$

Equation 14 can also be rewritten as:

$$a'x_n + b - \log[(1-p_n)/p_n] > 0 .$$

Equation 15 can also be rewritten as:

$$a'x_n + b - \log[(1-p_n)/p_n] < 0 .$$

If the previous conditions are not meet, determinator 205 declares the frame unvoiced. Equations 14 and 15 represent decision regions for making the voicing decision. The log term of the rewritten forms of equations 14 and 15 can be eliminated with some change of performance. Advantageously, in the present example, the element corresponding to power is the log of the speech energy.

Generator 206 is responsive to the information received via path 214 from calculator 204 to calculate the distance measure, A, as follows. First, the discriminant variable, d, is calculated by equation 16 as follows:

$$d = a'x_n + b - \log[(1-p_n)/p_n] . \qquad (16)$$

Advantageously, it would be obvious to one skilled in the art to use different types of voicing detectors to generate a value similar to d for use in the following equations. One such detector would be an auto-correlation detector. If the frame is voiced, the equations 17 through 20 are solved as follows:

$$m_1 = (1-z)\, m_1 + zd\,, \quad (17)$$

$$s_1 = (1-z)\, s_1 + zd^2, \text{ and} \quad (18)$$

$$k_1 = s_1 - m_1^2 \quad (19)$$

where m_1 is the mean for voiced frames and k_1 is the variance for voiced frames.

The probability, P_d, that determinator 205 will declare a frame unvoiced is calculated by the following equation:

$$P_d = (1-z)\, P_d\,. \quad (20)$$

Advantageously, P_d is initially set to .5.

If the frame is unvoiced, equations 21 through 24 are solved as follows:

$$m_0 = (1-z)\, m_0 + zd\,, \quad (21)$$

$$s_0 = (1-z)s_0 + zd^2, \text{and} \quad (22)$$

$$k_0 = s_0 - m_0^2\,. \quad (23)$$

The probability, P_d, that determinator 205 will declare a frame unvoiced is calculated by the following equation:

$$P_d = (1-z)\, P_d + z\,. \quad (24)$$

After calculating equation 16 through 22 the distance measure or merit value is calculated as follows:

$$A^2 = \frac{P_d\,(1 - P_d)\,(m_1 - m_0)^2}{(1 - P_d)k_1 + P_d k_0} \quad (25)$$

Equation 25 uses Hotelling's two-sample T^2 statistic to calculate the distance measure. For equation 25, the larger the merit value the greater the separation. However, other merit values exist where the smaller the merit value the greater the separation. Advantageously, the distance measure can also be the Mahalanobis distance which is given in the following equation:

$$A^2 = \frac{(m_1 - m_0)^2}{(1 - P_d)k_1 + P_d k_0}\,. \quad (26)$$

Advantageously, a third technique is given in the following equation:

$$A^2 = 2\frac{(m_1 - m_0)^2}{(k_1 + k_0)} . \qquad (27)$$

Advantageously, a fourth technique for calculating the distance measure is illustrated in the following equation:

$$A^2 = a'(v_n - u_n) \qquad (28)$$

Discriminant detector 102 makes the unvoiced/voiced decision by transmitting information to multiplexer 105 via path 107 indicating a voiced frame if a'x + b > 0. If this condition is not true, then detector 102 indicates an unvoiced frame. The values for vector a and scalar b used by detector 102 are advantageously identical to the initial values of a and b for statistical voiced detector 103.

Detector 102 determines the distance measure in a manner similar to generator 206 by performing calculations similar to those given in equations 16 through 28.

In flow chart form, FIGS. 3 and 4 illustrate, in greater detail, the operations performed by statistical voiced detector 103 of FIG.2. Blocks 302 and 300 implement blocks 202 and 201 of FIG. 2, respectively. Blocks 304 through 318 implement statistical calculator 203. Blocks 320 and 322 implement weights calculator 204, and blocks 326 through 338 implement block 205 of FIG.2. Generator 206 of FIG. 2 is implemented by block 340. Subtractor 207 is implemented by block 308 or block 324.

Block 302 calculates the vector which represents the average of the classifiers for the present frame and all previous frames. Block 300 determines whether speech or silence is present in the present frame; and if silence is present in the present frame, the mean for each classifier is subtracted from each classifier by block 324 before control is transferred to decision block 326. However, if speech is present in the present frame, then the statistical and weights calculations are performed by blocks 304 through 322. First, the average vector is found in block 302. Second, the sums of the squares and products matrix is calculated in block 304. The latter matrix along with the vector X representing the mean of the classifiers for the present and past frames is then utilized to calculate the covariance matrix, T, in block 306. The mean X is then subtracted from the classifier vector x_n in block 308.

Block 310 then calculates the probability that the present frame is unvoiced by utilizing the present weight vector a, the present threshold value b, and the classifier vector for the present frame, x_n. After calculating the probability that the present frame is unvoiced, the probability that the present frame is voiced is calculated by block 312. Then, the overall probability, p_n, that any frame will be unvoiced is calculated by block 314.

Blocks 316 and 318 calculate two vectors: in and v. The values contained in vector u represent the statistical average values that each classifier would have if the frame were unvoiced. Whereas, vector v contains values representing the statistical average values that each classifier would have if the frame were voiced. The actual vectors of classifiers for the present and previous frames are clustered around either vector u or vector v. The vectors representing the classifiers for the previous and present frames are clustered around vector u if these frames are found to be unvoiced; otherwise, the previous classifier vectors are clustered around vector v.

After execution of blocks 316 and 318, control is transferred to decision block 320. If N is greater than 99, control is transferred to block 322; otherwise, control is transferred to block 326. Upon receiving control, block 322 then calculates a new weight vector a and a new threshold value b. The vector a and value b are used in the next sequential frame by the preceding blocks in FIG. 3. Advantageously, if N is required to be greater than infinity, vector a and scalar b will never be changed, and detector 103 will adapt solely in response to vectors v and in as illustrated in blocks 326 through 338.

Blocks 326 through 338 implement u/v determinator 205 of FIG. 2. Block 326 determines whether the power term of vector v of the present frame is greater than or equal to the power term of vector u. If this condition is true, then decision block 328 is executed. The latter decision block determines whether the test for voiced or unvoiced is met. If the frame is found to be voiced in decision block 328, then the frame is so marked as voiced by block 330 otherwise the frame is marked as unvoiced by block 332. If the power term of vector v is less than the power term of vector in for the present frame, blocks 334 through 338 function are executed and function in a similar manner. Finally, block 340 calculates the distance measure.

In flow chart form, FIG. 5 illustrates, in greater detail the operations performed by block 340 of FIG. 4. Decision block 501 determines whether the frame has been indicated as unvoiced or voiced by examining the calculations 330, 332, 336, or 338. If the frame has been designated as voiced, path 507 is selected.

Block 510 calculates probability P_d, and block 502 recalculates the mean, m_1, for the voiced frames and block 503 recalculates the variance; k_1, for voiced frames. If the frame was determined to be unvoiced, decision block 501 selects path 508. Block 509 recalculates probability P_d, and block 504 recalculates mean, m_0, for unvoiced frames, and block 505 recalculates the variance k_0 for unvoiced frames. Finally, block 506 calculates the distance measure by performing the calculations indicated.

Claims

1. An apparatus analyzing silence-free speech for making a voicing decision in speech signals, comprising:

   means responsive to said speech signals for sampling said speech signals to produce digital speech signals, to form frames of said digital non-training set speech signals, and to process each frame to generate a set of classifiers defining speech attributes;

   CHARACTERISED BY

   first means (203) responsive to said set of classifiers defining speech attributes of one of said frames of digital speech for calculating a set of parameters defining a set of statistical distributions;

   second means (204) responsive to the calculated set of statistical distributions based on said one of said frames of digital speech for calculating a set of weights each associated with one of said classifiers;

   third means (205) responsive to the calculated set of weights and classifiers and said set of statistical distributions for making the voicing decision for said frame of speech;

   said second means comprising means (322) for calculating a threshold value in response to said set of said parameters;

   means (213,204) for communicating said set of said weights and said threshold value to said first means to be used for calculating another set of parameters for another one of said frames of speech; and

   said first means being further responsive to the communicated set of weights and another set of classifiers defining said speech attributes of said other one of said frames for calculating another set of statistical parameters defining another updated set of statistical distributions.

2. The apparatus of claim 1 wherein said first means comprises means for calculating the average of each of said classifiers over previous ones of said speech frames; and

   means responsive to said average ones of said classifiers for said previous ones of said speech frames and said communicated set of weights and said other set of classifiers for determining said other set of statistical parameters.

3. The apparatus of claim 2 wherein said first means further comprises means for calculating the probability that said other set of classifiers represents an unvoiced frame and the probability that said other set of classifiers represents a voiced frame; and

   means for calculating the overall probability that any frame is unvoiced.

4. The apparatus of claim 3 wherein said first means further comprises means for calculating a set of statistical average classifiers representing an unvoiced frame and a set of statistical average classifiers representing a voiced frame.

5. The apparatus of claim 4 wherein said first means further comprises means for calculating a covariance matrix from said set of averaged classifiers representing an unvoiced frame for said other one of saids frames and said set of classifiers representing an unvoiced frame for said other one of said frames.

6. The apparatus of claim 5 wherein said second means responsive to the covariance matrix and said sets of statistical average classifiers for both voiced and unvoiced frames and said overall probability for a frame being unvoiced for determining said other set of parameters.

7. The apparatus of claim 6 wherein said third means responsive to said other set of parameters and said sets of statistical average classifiers for unvoiced and voiced frames for determining the presence of said fundamental frequency in said other one of said frames.

Patentansprüche

Block 510 calculates probability P_d, and block 502 recalculates the mean, m_1, for the voiced frames and block 503 recalculates the variance, k_1, for voiced frames. If the frame was determined to be unvoiced, decision block 501 selects path 508. Block 509 recalculates probability P_d, and block 504 recalculates mean, m_0, for unvoiced frames, and block 505 recalculates the variance k_0 for unvoiced frames. Finally, block 506 calculates the distance measure by performing the calculations indicated.

Claims

1. An apparatus analyzing silence-free speech for making a voicing decision in speech signals, comprising:

   means responsive to said speech signals for sampling said speech signals to produce digital speech signals, to form frames of said digital non-training set speech signals, and to process each frame to generate a set of classifiers defining speech attributes;

   CHARACTERISED BY

   first means (203) responsive to said set of classifiers defining speech attributes of one of said frames of digital speech for calculating a set of parameters defining a set of statistical distributions;

   second means (204) responsive to the calculated set of statistical distributions based on said one of said frames of digital speech for calculating a set of weights each associated with one of said classifiers;

   third means (205) responsive to the calculated set of weights and classifiers and said set of statistical distributions for making the voicing decision for said frame of speech;

   said second means comprising means (322) for calculating a threshold value in response to said set of said parameters;

   means (213,204) for communicating said set of said weights and said threshold value to said first means to be used for calculating another set of parameters for another one of said frames of speech; and

   said first means being further responsive to the communicated set of weights and another set of classifiers defining said speech attributes of said other one of said frames for calculating another set of statistical parameters defining another updated set of statistical distributions.

2. The apparatus of claim 1 wherein said first means comprises means for calculating the average of each of said classifiers over previous ones of said speech frames; and

   means responsive to said average ones of said classifiers for said previous ones of said speech frames and said communicated set of weights and said other set of classifiers for determining said other set of statistical parameters.

3. The apparatus of claim 2 wherein said first means further comprises means for calculating the probability that said other set of classifiers represents an unvoiced frame and the probability that said other set of classifiers represents a voiced frame; and

   means for calculating the overall probability that any frame is unvoiced.

4. The apparatus of claim 3 wherein said first means further comprises means for calculating a set of statistical average classifiers representing an unvoiced frame and a set of statistical average classifiers representing a voiced frame.

5. The apparatus of claim 4 wherein said first means further comprises means for calculating a covariance matrix from said set of averaged classifiers representing an unvoiced frame for said other one of saids frames and said set of classifiers representing an unvoiced frame for said other one of said frames.

6. The apparatus of claim 5 wherein said second means responsive to the covariance matrix and said sets of statistical average classifiers for both voiced and unvoiced frames and said overall probability for a frame being unvoiced for determining said other set of parameters.

7. The apparatus of claim 6 wherein said third means responsive to said other set of parameters and said sets of statistical average classifiers for unvoiced and voiced frames for determining the presence of said fundamental frequency in said other one of said frames.

Patentansprüche

1. Vorrichtung zur Bereitstellung einer Stimmhaft-Stimmlos-Entscheidung für Sprachsignale bei der Analyse ruhefreier Sprache mit
einer Einrichtung, die unter Ansprechen auf die Sprachsignale diese zur Erzeugung digitaler Sprachsignale abtastet, Rahmen der digitalen Sprachsignale bildet und jeden Rahmen zur Erzeugung eines Satzes von Sprachattribute definierenden Klassifizierern verarbeitet, dadurch gekennzeichnet,
daß eine erste Einrichtung (203) vorgesehen ist, die unter Ansprechen auf den Sprachattribute definierenden Satz von Klassifizierern eines Rahmens digitaler Sprache einen Satz von Parametern berechnet, die einen Satz statistischer Verteilungen definieren, ferner
eine zweite Einrichtung (204), die unter Ansprechen auf den berechneten Satz statistischer Verteilungen auf der Grundlage des einen Rahmens digitaler Sprache einen Satz von Bewertungen berechnet, die je einem der Klassifizierer zugeordnet sind, und
eine dritte Einrichtung (205), die unter Ansprechen auf den berechneten Satz von Bewertungen und Klassifizierern und den Satz statistischer Verteilungen die Stimmhaft-Stimmlos-Entscheidung für den Sprachrahmen trifft,
daß die zweite Einrichtung eine Einrichtung (322) zur Berechnung eines Schwellenwertes unter Ansprechen auf den Satz von Parametern und eine Einrichtung (213, 204) zur Übertragung des Satzes von Bewertungen und des Schwellenwertes zu der ersten Einrichtung aufweist, um für die Berechnung eines weiteren Satzes von Parametern für einen weiteren Sprachrahmen benutzt zu werden, und
daß die erste Einrichtung ferner auf den übertragenen Satz von Bewertungen und einen weiteren Satz von Klassifizierern, die die Sprachattribute des weiteren Rahmens definieren, um einen weiteren Satz statistischer Parameter zu berechnen, der einen weiteren, aktualisierten Satz statistischer Verteilungen definiert.

2. Vorrichtung nach Anspruch 1,
bei der die erste Einrichtung eine Einrichtung zur Berechnung des Mittelwertes jedes Klassifizierers über vorhergehende Sprachrahmen und eine Einrichtung umfaßt, die unter Ansprechen auf den Mittelwert der Klassifizierer für die vorhergehenden Sprachrahmen und den übertragenen Satz von Bewertungen sowie den weiteren Satz von Klassifizierern den weiteren Satz statistischer Parameter bestimmt.

3. Vorrichtung nach Anspruch 2,
bei der die erste Einrichtung ferner eine Einrichtung zur Berechnung der Wahrscheinlichkeit aufweist, daß der weitere Satz von Klassifizierern einen stimmlosen Rahmen darstellt und der Wahrscheinlichkeit, daß der weitere Satz von Klassifizierern einen stimmhaften Rahmen darstellt, und eine Einrichtung zur Berechnung der Gesamtwahrscheinlichkeit, daß jeder Rahmen stimmlos ist.

4. Vorrichtung nach Anspruch 3,
bei der die erste Einrichtung ferner eine Einrichtung zur Berechnung eines Satzes statistischer Mittelwertklassifizierer aufweist, die einen stimmlosen Rahmen darstellen, und eines statistischer Mittelwertklassifizierer, die einen stimmhaften Rahmen darstellen.

5. Vorrichtung nach Anspruch 4,
bei der die erste Einrichtung ferner eine Einrichtung zur Berechnung einer Covarianz-Matrix aus dem Satz von einen stimmlosen Rahmen darstellenden Mittelwertklassifizierern für den weiteren Rahmen und aus dem Satz von einen stimmlosen Rahmen darstellenden Klassifizierern für den weiteren Rahmen.

6. Vorrichtung nach Anspruch 5,
bei der die zweite Einrichtung unter Ansprechen auf die Covarianz-Matrix und die Sätze statistischer Mittelwertklassifizierer für sowohl stimmhafte als auch stimmlose Rahmen und die Gesamtwahrscheinlichkeit dafür, daß ein Rahmen stimmlos ist, den weiteren Satz von Parametern bestimmt.

7. Vorrichtung nach Anspruch 6,
bei der die dritte Einrichtung unter Ansprechen auf den weiteren Satz von Parametern und die Sätze von statistischen Mittelwertklassifizierern für stimmlose und stimmhafte Rahmen das Vorhandensein der Grundfrequenz in dem weiteren Rahmen bestimmt.

* C L M S *

Revendications

1. Un appareil d'analyse de parole ne contenant pas d'intervalles de silence, destiné à prendre une décision de voisement concernant des signaux de parole, comprenant :

des moyens qui réagissent aux signaux de parole en échantillonnant ces signaux de parole de façon à produire des signaux de parole numériques, pour former des trames de ces signaux de parole numériques, et pour traiter chaque trame de façon à générer un ensemble de classificateurs définissant des attributs de la parole;

CARACTERISE PAR

des premiers moyens (203) qui réagissent à l'ensemble de classificateurs définissant des attributs de la parole de l'une des trames de parole numérique, en calculant un ensemble de paramètres définissant un ensemble de distributions statistiques;

des seconds moyens (204) qui réagissent à l'ensemble de distributions statistiques calculé, base sur l'une des trames de parole numérique, en calculant un ensemble de poids, avec chaque poids associé à l'un des classificateurs;

des troisièmes moyens (205) qui réagissent à l'ensemble de poids calculés et aux classificateurs et à l'ensemble de distributions statistiques, en prenant la décision de voisement pour la trame de parole considérée;

les seconds moyens comprenant des moyens (322) qui sont destinés à calculer une valeur de seuil sous la dépendance de l'ensemble des paramètres;

des moyens (213, 204) qui sont destinés à communiquer l'ensemble de poids et la valeur de seuil aux premiers moyens, pour l'utilisation dans le calcul d'un autre ensemble de paramètres pour une autre des trames de parole; et

les premiers moyens réagissant en outre à l'ensemble de poids communiqué et à un autre ensemble de classificateurs définissant les attributs de la parole de l'autre trame, en calculant un autre ensemble de paramètres statistiques définissant un autre ensemble actualisé de distributions statistiques.

2. L'appareil de la revendication 1, dans lequel les premiers moyens comprennent des moyens qui sont destinés à calculer la moyenne de chacun des classificateurs sur des trames de parole précédentes; et

des moyens qui réagissent à la moyenne des classificateurs pour les trames de parole précédentes et à l'ensemble de poids communiqué, ainsi qu'à l'autre ensemble de classificateurs, en déterminant l'autre ensemble de paramètres statistiques.

3. L'appareil de la revendication 2, dans lequel les premiers moyens comprennent en outre des moyens qui sont destinés à calculer la probabilité que l'autre ensemble de classificateurs représente une trame non voisée, et la probabilité que l'autre ensemble de classificateurs représente une trame voisée; et

des moyens destinés à calculer la probabilité globale d'une trame quelconque soit non voisée.

4. L'appareil de la revendication 3, dans lequel les premiers moyens comprennent en outre des moyens qui sont destinés à calculer un ensemble de classificateurs moyens statistiques représentant une trame non voisée, et un ensemble de classificateurs moyens statistiques représentant une trame voisée.

5. L'appareil de la revendication 4, dans lequel les premiers moyens comprennent en outre des moyens qui sont destinés à calculer une matrice de covariance à partir de l'ensemble de classificateurs moyens représentant une trame non voisée pour l'autre trame parmi les trames précitées, et de l'ensemble de classificateurs représentant une trame non voisée pour l'autre trame parmi les trames précitées.

6. L'appareil de la revendication 5, dans lequel les seconds moyens réagissent à la matrice de covariance et aux ensembles précités de classificateurs moyens statistiques pour des trames voisées et non voisées, ainsi qu'à la probabilité globale qu'une trame soit non voisée, en déterminant l'autre ensemble de paramètres.

7. L'appareil de la revendication 6, dans lequel les troisièmes moyens réagissent à l'autre ensemble de paramètres et aux ensembles de classificateurs moyens statistiques pour des trames voisées et non voisées en déterminant la présence de la fréquence fondamentale dans l'autre trame parmi les trames précitées.