PCT WORLD INTELLECTUAL PROPE
International Bu

WO 9608005A1

INTERNATIONAL APPLICATION PUBLISHED UNDER

(51) International Patent Classification 6 : G10L 5/00 | A1

(11) International Publication Number: WO 96/08005

(43) International Publication Date: 14 March 1996 (14.03.96)

(21) International Application Number: PCT/US95/08869

(22) International Filing Date: 14 July 1995 (14.07.95)

(30) Priority Data:
08/302,067 7 September 1994 (07.09.94) US

(71) Applicant: MOTOROLA INC. [US/US]; 1303 East Algonquin Road, Schaumburg, IL 60196 (US).

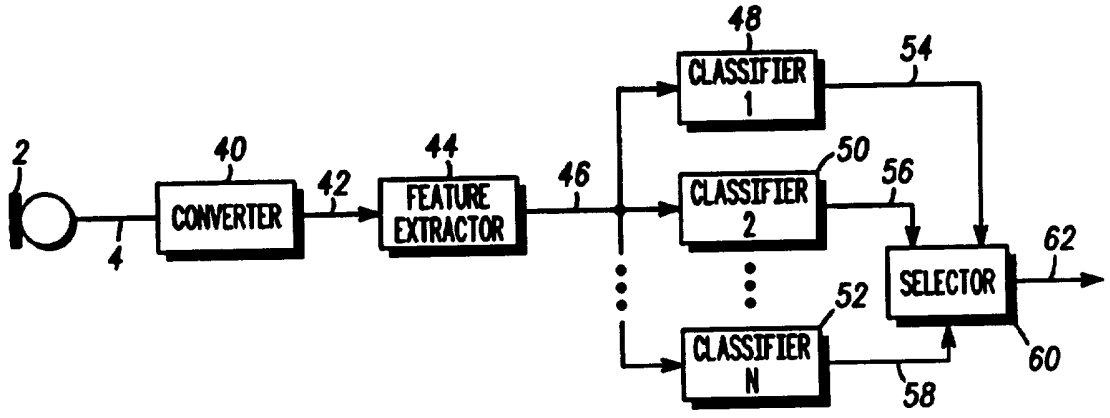
(72) Inventor: WANG, Shay-Ping, T.; 1701 E. Edgewood Lane, Long Grove, IL 60047 (US).

(74) Agents: STUCKMAN, Bruce, E. et al.; Motorola Inc., Intellectual Property Dept., 1303 East Algonquin Road, Schaumburg, IL 60196 (US).

(81) Designated States: AM, AT, AU, BB, BG, BR, BY, CA, CH, CN, CZ, DE, DK, EE, ES, FI, GB, GE, HU, IS, JP, KE, KG, KP, KR, KZ, LK, LR, LT, LU, LV, MD, MG, MN, MW, MX, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, TJ, TM, TT, UA, UG, UZ, VN, European patent (AT, BE, CH, DE, DK, ES, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, ML, MR, NE, SN, TD, TG), ARIPO patent (KE, MW, SD, SZ, UG).

Published
With international search report.

(54) Title: SYSTEM FOR RECOGNIZING SPOKEN SOUNDS FROM CONTINUOUS SPEECH AND METHOD OF USING SAME



(57) Abstract

A system for recognizing spoken sounds from continuous speech includes a plurality of classifiers (48) - (52) and a selector. Each of the classifiers implements a discriminant function (60) which is based on a polynomial expansion. By determining the polynomial coefficients of a discriminant function, the corresponding classifier is tuned to classify a specific spoken sound. The selector (60) utilizes the classifier outputs (54) - (58) to identify the spoken sounds. A method of using the system is also provided.

FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AT	Austria	**GB**	United Kingdom	**MR**	Mauritania
AU	Australia	**GE**	Georgia	**MW**	Malawi
BB	Barbados	**GN**	Guinea	**NE**	Niger
BE	Belgium	**GR**	Greece	**NL**	Netherlands
BF	Burkina Faso	**HU**	Hungary	**NO**	Norway
BG	Bulgaria	**IE**	Ireland	**NZ**	New Zealand
BJ	Benin	**IT**	Italy	**PL**	Poland
BR	Brazil	**JP**	Japan	**PT**	Portugal
BY	Belarus	**KE**	Kenya	**RO**	Romania
CA	Canada	**KG**	Kyrgystan	**RU**	Russian Federation
CF	Central African Republic	**KP**	Democratic People's Republic of Korea	**SD**	Sudan
CG	Congo			**SE**	Sweden
CH	Switzerland	**KR**	Republic of Korea	**SI**	Slovenia
CI	Côte d'Ivoire	**KZ**	Kazakhstan	**SK**	Slovakia
CM	Cameroon	**LI**	Liechtenstein	**SN**	Senegal
CN	China	**LK**	Sri Lanka	**TD**	Chad
CS	Czechoslovakia	**LU**	Luxembourg	**TG**	Togo
CZ	Czech Republic	**LV**	Latvia	**TJ**	Tajikistan
DE	Germany	**MC**	Monaco	**TT**	Trinidad and Tobago
DK	Denmark	**MD**	Republic of Moldova	**UA**	Ukraine
ES	Spain	**MG**	Madagascar	**US**	United States of America
FI	Finland	**ML**	Mali	**UZ**	Uzbekistan
FR	France	**MN**	Mongolia	**VN**	Viet Nam
GA	Gabon				

SYSTEM FOR RECOGNIZING SPOKEN SOUNDS FROM CONTINUOUS SPEECH AND METHOD OF USING SAME

Related Inventions

The present invention is related to the following invention which is assigned to the same assignee as the present invention:

(1) "Neural Network and Method of Using Same", having Serial No. 08/076,601, filed on June 14, 1993.

(2) "Speech-Recognition System Utilizing Neural Networks and Method of Using Same", having Serial No. ___,___, filed on _____.

The subject matter of the above-identified related invention is hereby incorporated by reference into the disclosure of this invention.

Technical Field

This invention relates generally to speech recognition systems and, in particular, to a speech recognition system which recognizes continuous speech.

Background of the Invention

For many years, scientists have been trying to find a means to simplify the interface between man and machine. Input devices such as the keyboard, mouse, touch screen, and pen are currently the most commonly used tools for implementing a man/machine interface. However, a simpler and more natural interface between man and machine may be human speech. A device which automatically recognizes speech would provide such an interface.

Applications for an automated speech-recognition device include a database query technique using voice commands, voice input for quality control in a manufacturing process, a voice-dial cellular phone which

would allow a driver to focus on the road while dialing, and a voice-operated prosthetic device for the physically disabled.

Unfortunately, automated speech recognition is not a trivial task. One reason is that speech tends to vary considerably from one person to another. For instance, the same word uttered by several persons may sound significantly different due to differences in accent, speaking speed, gender, or age. In addition to speaker variability, co-articulation effects, speaking modes (shout/whisper), and background noise present enormous problems to speech-recognition devices.

Since the late 1960's, various methodologies have been introduced for automated speech recognition. While some methods are based on extended knowledge with corresponding heuristic strategies, others rely on speech databases and learning methodologies. The latter methods include dynamic time-warping (DTW) and hidden-Markov modeling (HMM). Both of these methods, as well as the use of time-delay neural networks (TDNN), are discussed below.

Dynamic time-warping is a technique which uses an optimization principle to minimize the errors between an unknown spoken word and a stored template of a known word. Reported data shows that the DTW technique is very robust and produces good recognition. However, the DTW technique is computationally intensive. Therefore, it is currently impractical to implement the DTW technique for real-world applications.

Instead of directly comparing an unknown spoken word to a template of a known word, the hidden-Markov modeling technique uses stochastic models for known words and compares the probability that the unknown word was generated by each model. When an unknown word is uttered, the HMM technique will check the sequence (or state) of the word, and find the model that provides the best match. The HMM technique has been successfully used in many commercial applications; however, the technique has many drawbacks.

These drawbacks include an inability to differentiate acoustically similar words, a susceptibility to noise, and computational intensiveness.

Recently, neural networks have been used for problems that are highly unstructured and otherwise intractable, such as speech recognition. A time-delay neural network is a type of neural network which addresses the temporal effects of speech by adopting limited neuron connections. For limited word recognition, a TDNN shows slightly better results than the HMM method. However, a TDNN suffers from some serious drawbacks.

First, the training time for a TDNN is very lengthy, on the order of several weeks. Second, the training algorithm for a TDNN often converges to a local minimum, which is not the globally optimum solution.

In summary, the drawbacks of existing known methods of automated speech-recognition (e.g. algorithms requiring impractical amounts of computation, limited tolerance to speaker variability and background noise, excessive training time, etc.) severely limit the acceptance and proliferation of speech-recognition devices in many potential areas of utility. There is thus a significant need for an automated speech-recognition system which provides a high level of accuracy, is immune to background noise, does not require repetitive training or complex computations, and is insensitive to differences in speakers.

Summary of Invention

It is therefore an advantage of the present invention that spoken sounds can be recognized from continuous speech and the recognition rate is insensitive to differences in speakers.

It is a further advantage of the present invention that spoken sounds can be recognized from continuous speech

wherein the recognition rate is not adversely affected by background noise.

Another advantage of the present invention is to provide both a speech-recognition method and system, neither of which requires repetitive training.

Yet another advantage of the present invention is to provide both a method and system for continuous speech recognition, either of which operates with a vast reduction in computational complexity.

These and other advantages are achieved in accordance with a preferred embodiment of the invention by providing a method of recognizing a spoken sound from continuously spoken speech. In this method, the continuously spoken speech includes a plurality of spoken sounds. The method is described by the following steps. A speech signal is generated from the continuously spoken speech. Next, the speech signal is processed to form a feature frame by extracting a plurality of features corresponding to the continuously spoken speech at an instant in time. The feature frame is distributed to a plurality of classifiers. In addition, each classifier implements a discriminant function to generate a classifier output signal in response to the feature frame. Finally, the spoken sound corresponding to the instant of time is identified by comparing the classifier output signals from each of the plurality of classifiers.

In another embodiment of the present invention there is provided a system for recognizing spoken sounds from continuously spoken speech. The system includes the following elements. First, a plurality of classifiers. Each of the classifiers receives a feature frame and implements a discriminant function to generate an output in response to the feature frame. The feature frame is derived from the continuously spoken speech. The second element, a selector, is responsive to the output of each of the classifiers. The selector identifies the spoken sound corresponding to an instant in time by comparing the

classifier output signals from each of the plurality of classifiers.

Brief Description of the Drawings

The invention is pointed out with particularity in the appended claims. However, other features of the invention will become more apparent and the invention will be best understood by referring to the following detailed description in conjunction with the accompanying drawings in which:

FIG. 1 shows a contextual diagram of a speech-recognition system.

FIG. 2 shows a block diagram of a system for recognizing spoken sounds from continuous speech, in accordance with one embodiment of the present invention.

FIG. 3 is a block diagram of a classifier ~~which is~~ in accordance with a preferred embodiment of the present invention.

FIG. 4 shows a flow diagram of a method of training a speech-recognition system to identify spoken sounds from continuous speech in accordance with the present invention.

FIG. 5 shows a flow diagram of a method of recognizing a spoken sound from continuous speech in accordance with a preferred embodiment of the present invention.

Detailed Description of a Preferred Embodiment

It will be understood by one of ordinary skill in the art that the methods of the present invention may be implemented in hardware or software, or any combination thereof, and that the terms, "continuous speech" and "continuously spoken speech" are used interchangeably in this description.

FIG. 1 shows a contextual block diagram of a speech-recognition system. The diagram shows microphone 2 or equivalent means for receiving audio input in the form of

speech input and converting sound into electrical energy. Speech recognition system 6 receives signals from microphone 2 over transmission medium 4 and performs various tasks such as waveform sampling, analog-to-digital (A/D) conversion, feature extraction and classification. Speech recognition system 6 provides the identity of spoken sounds to computer 10 via bus 8. Computer 10 executes commands or programs which may utilize the data provided by speech recognition system 6.

One of ordinary skill in the art will understand that speech recognition system 6 may transmit spoken sound identities to devices other than a computer. For example, a communication network, data storage system, or transcription device could be substituted for computer 10.

FIG. 2 shows a block diagram of a system for recognizing spoken sounds from continuous speech, in accordance with one embodiment of the present invention. The system comprises microphone 2, converter 40, feature extractor 44, a plurality of classifiers, and selector 60 which elements, except for microphone 2, comprise speech recognition system 6 (FIG 2). In the example given by FIG. 2, three classifiers are shown; these are classifiers 48, 50, and 52.

Continuous speech is received by microphone 2 and converted to signals which are transmitted across transmission medium 4 to converter 40. Converter 40 performs various functions which utilize the speech signals. These functions include waveform sampling and analog-to-digital (A/D) conversion. Converter 40 generates as output a speech signal which is passed to feature extractor 44 via bus 42. Feature extractor 44 creates a set of features, or measurements, which contain much of the same information as the speech signal, but with reduced dimensionality. These features are distributed by bus 46 to a plurality of classifiers, of which classifiers 48, 50, and 52 are shown. Generally, each classifier implements a discriminant function which determines whether a set of

features belongs to a particular class. The result of computing each discriminant function, referred to as a classifier output signal, is sent to selector 60. In the example given, classifiers 48, 50, and 52 transfer classifier output signals across buses 54, 56, and 58, respectively, to selector 60. Selector 60 compares the classifier output signals with one another, and based on the results of the comparison, selector 60 provides the identity of the spoken sound on output 62.

The operation of the system commences when a user speaks into microphone 2. In a preferred embodiment of the present invention, the system depicted by FIG. 2 is used for recognizing spoken sound from continuously spoken speech. Continuously spoken speech, or continuous speech, takes place when a person speaking into the microphone does not un-naturally pause between each spoken sound. Rather, the person speaking only pauses when the natural form of speech dictates a pause, such as at the end of a sentence. For this reason, continuous speech can be thought of as "natural" speech which occurs in an ordinary conversation. Continuously spoken speech includes at least one spoken sound, wherein a spoken sound may be a word, character, or phoneme. A phoneme is the smallest element of speech sound which indicates a difference in meaning. A character includes one or more phonemes, and a word includes one or more characters.

When a user utters continuous speech, microphone 2 generates a signal which represents the acoustic waveform of the speech. Typically, the signal from microphone 2 is an analog signal. This signal is then fed to converter 40 for digitization. Converter 40 includes appropriate means for A/D conversion as is generally well known to one of ordinary skill in the art. Converter 40 may use an A/D converter (not shown) to sample the signal from microphone 2 several thousand times per second (e.g. between 8000 and 14,000 times per second in a preferred embodiment of the present invention depending on the frequency components of

the speech signal from the microphone). Each of the samples is then converted to a digital word, wherein the length of the word is between 12 and 32 bits.

Those of ordinary skill in the art will understand that the sampling rate and word length of A/D converters may vary and that the numbers given above do not place any limitations on the sampling rate or word length of the A/D converter which is included in an embodiment of the present invention.

The speech signal from converter 40 comprises one or more of these digital words, wherein each digital word represents a sample of the continuous speech taken at an instant in time. The speech signal is passed to feature extractor 44 where the digital words, over an interval of time, are grouped into a data frame. In a preferred embodiment of the present invention each data frame represents 10 milliseconds of speech signal. However, one of ordinary skill in the art will recognize that other data frame durations may be used, depending on a number of factors such as the duration of the spoken sounds to be identified. The data frames are in turn subjected to cepstral analysis, a method of feature extraction, which is performed by feature extractor 44.

The cepstral analysis, or feature extraction, which is performed on the speech signal, results in a representation of the speech signal which characterizes the relevant features of the continuous speech over the interval of time. It can be regarded as a data reduction procedure that retains vital characteristics of the speech signal and eliminates undesirable interference from irrelevant characteristics of the speech signal, thus easing the decision-making process of the plurality of classifiers.

The cepstral analysis is performed as follows. First, a p-th order (typically p = 12 to 14) linear prediction analysis is applied to a set of digital words from the speech signal to yield p prediction coefficients. The

prediction coefficients are then converted into cepstrum coefficients using the following recursion formula:

$$c(n) = a(n) + \sum_{k=1}^{n-1} (1 - k/n)\ a(k)\ c(n - k) \qquad \text{Equation (1)}$$

wherein $c(n)$ represents the n^{th} cepstrum coefficient, $a(n)$ represents the n^{th} prediction coefficient, $1 \leq n \leq p$, p is equal to the number of cepstrum coefficients, n represents an integer index, and k represents an integer index, and $a(k)$ represents the k^{th} prediction coefficient and $c(n - k)$ represents the $(n - k)^{th}$ cepstrum coefficient.

The vector of cepstrum coefficients is usually weighted by a sine window of the form,

$$\alpha(n) = 1 + (L/2)\ \sin(\pi n/L) \qquad \text{Equation (2)}$$

wherein $1 \leq n \leq p$, and L is an integer constant, giving the weighted cepstrum vector, $C(n)$, wherein

$$C(n) = c(n)\ \alpha(n) \qquad \text{Equation (3)}$$

This weighting is commonly referred to as cepstrum liftering. The effect of this liftering process is to smooth the spectral peaks in the spectrum of the speech signal. It has also been found that cepstrum liftering suppresses the existing variations in the high and low cepstrum coefficients, and thus considerably improves the performance of the speech-recognition system.

The result of the cepstral analysis is a smoothed log spectra which corresponds to the frequency components of the speech signal over an interval of time. The significant features of the speech signal are thus preserved in the spectra. Feature extractor 44 generates a respective feature frame which comprises data points from the spectrum generated from a corresponding data frame.

The feature frame is then passed or distributed to the plurality of classifiers.

In a preferred embodiment of the present invention, a feature frame contains twelve data points, wherein each of the data points represents the value of cepstrally-smoothed spectrum at a specific frequency over the interval of time. The data points are 32-bit digital words. Those skilled in the art will understand that the present invention places no limits on the number of data points per feature frame or the bit length of the data points; the number of data points contained in a feature frame may be twelve or any other appropriate value, while the data point bit length may be 32 bits, 16 bits, or any other value.

In general, a classifier makes a decision as to which class an input pattern belongs. In a preferred embodiment of the present invention, each class is labeled with a spoken sound, and examples of the spoken sound are obtained from a predefined set of spoken sounds (the training set) and used to determine boundaries between the classes, boundaries which maximize the recognition performance for each class.

Upon receiving a feature frame, each of the classifiers 48, 50,...52 employ a parametric decision method to determine whether a feature frame belongs to a certain class. With this method, each classifier computes a different discriminant function $y_j(X)$, wherein $X = \{x_1, x_2, \ldots, x_i\}$ is the set of data points contained in a feature frame, i is an integer index, and j is an integer index corresponding to the classifier. Upon receiving a feature frame, the classifiers compute their respective discriminant functions and provide the results of their computations as classifier output signals. Generally, the magnitude of a classifier output signal indicates whether a feature frame belongs to the class which corresponds to the discriminant function. In a preferred embodiment of the present invention, the magnitude of a classifier output

signal is directly proportional to the likelihood that the feature frame belongs to the corresponding class.

The discriminant functions computed by the classifiers are based upon the use of a polynomial expansion and, in a loose sense, the use of an orthogonal function, such as a sine, cosine, exponential/logarithmic, Fourier transformation, Legendre polynomial, non-linear basis function such as a Volterra function or a radial basis function, or the like, or a combination of polynomial expansion and orthogonal functions.

A preferred embodiment of the present invention employs a polynomial expansion of which the general case is represented by Equation 4 as follows:

$$y = \sum_{i=1}^{m} w_{i-1} \; x_1^{g_{1i}} \; x_2^{g_{2i}} \; . \; . \; . \; x_n^{g_{ni}} \qquad \text{Equation (4)}$$

where x_i represent the classifier inputs and can be a function such as $x_i = f_i(z_j)$, wherein z_j is any arbitrary variable, and where the indices i, j, and m may be any integers; where y represents the output of the classifier; where w_{i-1} represent the coefficient for the ith term; where $g_{1i}, \; . \; . \; ., \; g_{ni}$ represent the exponents for the ith term and are integers; and n is the number of classifier inputs.

In the example shown by FIG. 2, the classifier output signal of classifier 48 is passed to selector 62 across bus 54; the classifier output signal of classifier 50 is passed across bus 56 to selector 60; and classifier output signal of character classifier 52 is passed across bus 58 to selector 60.

Selector 60 determines which of the classifier output signals has the largest magnitude and then produces a representation of the corresponding spoken sound identity on output 62. In one embodiment of the present invention,

the representation produced by selector 60 is digital word coded in a computer-readable format. However, one of ordinary skill in the art will appreciate that the representation provided on output 62 may vary in form depending on the application of the system. For example, output 62, as with any of the signals herein described, could be an analog or optical signal.

In one embodiment of the present invention, the system shown in FIG. 2 is implemented by software running on a processor such as a microprocessor. However, one of ordinary skill will recognize that a programmable logic array, ASIC, or other digital logic device could also be used to implement the functions performed by the system shown in FIG 2.

FIG. 3 is a block diagram of a classifier which is in accordance with a preferred embodiment of the present invention. Classifier 110 is a possible implementation of one of the plurality of classifiers depicted in FIG. 2. Classifier 110 includes a plurality of computing elements, of which computing element 111, 113, and 115 are shown. Classifier 110 also includes summation circuit 117.

A polynomial expansion is calculated by classifier 110 in the following manner. A plurality of data inputs x_1, x_2, . . ., x_n are fed into classifier 110 using bus 119 and then distributed to the plurality of computing elements, represented by 111, 113, and 115. Typically, the data inputs would be data points from a feature frame. Each computing element determines which of the data inputs to receive and computes one or more terms in the polynomial expansion. After computing a term, a computing element passes the term to summing circuit 117 which sums the terms computed by the computing elements and places the sum on output 133.

For example, FIG. 3 depicts the computation of the polynomial $y = x_1 g_{11}\, x_2 g_{21} + x_1 g_{12}\, x_2 g_{22} + \ldots x_n g_{nm}$. Computing element 111 computes the term $x_1 g_{11}\, x_2 g_{21}$ and then sends it to summing circuit 117 over bus 127; computing

element 113 computes the term $x_1 g_{12}$ $x_2 g_{22}$ and then sends it to summing circuit 117 over bus 129; and computing element 115 computes the term $x_n g_{nm}$ and then sends it to summing circuit 117 over bus 131. Upon receiving the terms from the computing elements, summing circuit 117 sums the terms and places the result of the polynomial expansion, y, on output 133.

It will be apparent to one of ordinary skill that classifier 110 is capable of computing polynomials of the form given by Equation 1 which have a number of terms different from the above example, and polynomials whose terms are composed of data inputs different from those of the above example.

In one embodiment of the present invention, classifier 110 is implemented by software running on a processor such as a microprocessor. However, one of ordinary skill in the art will recognize that a programmable logic array, ASIC or other digital logic device could also be used to implement the functions performed by the classifier 110.

FIG. 4 shows a flow diagram of a method of training a speech-recognition system to identify spoken sounds from continuous speech. A speech-recognition system constructed in accordance with an embodiment of present invention has principally two modes of operation: (1) a training mode in which examples of spoken sounds are used to train the plurality of classifiers, and (2) a recognition mode in which spoken sounds in continuous speech are identified. Referring to FIG. 2, generally, a user must train the plurality of classifiers by providing examples of all of the spoken sounds that the system is to recognize.

In an embodiment of the present invention, a classifier may be trained by tuning the coefficients of a discriminant function which is based on a polynomial expansion of the form given by Equation 4. For the discriminant function to effectively classify input data, the coefficient, w_{i-1}, of each term in the polynomial

expansion must be determined. This can be accomplished by the use of the following training method.

In box 140, a plurality of spoken sound examples is provided. A spoken sound example comprises two components. The first component is a set of samples of the spoken sound, and the second component is a corresponding desired classifier output signal.

Next, in box 142, the trainer compares the number of spoken sound examples with the number of polynomial coefficients in the discriminate function.

In decision box 144, a check is made to determine whether the number of coefficients equal to the number of spoken sound examples. If so, the method proceeds to box 146. If not, the method proceeds to box 148.

In box 146, a matrix inversion technique is used to solve for the value of each polynomial coefficient.

In box 148, a least squares estimation technique is used to solve for the value of each polynomial coefficient. Suitable least-squares estimation techniques include, for example, least-squares, extended least-squares, pseudo-inverse, Kalman filter, maximum-likelihood algorithm, Bayesian estimation, and the like.

In implementing a classifier which is usable in an embodiment of the present invention, one generally selects the number of computing elements in the classifier to be equal to or less than the number of examples presented to the learning machine.

FIG. 5 shows a flow diagram of a method of recognizing a spoken sound from continuous speech in accordance with a preferred embodiment of the present invention. In box 150, a speech signal is generated from continuously spoken speech.

Next, in box 152, a plurality of features are extracted from the speech signal. The features correspond to the continuously spoken speech over an interval of time. In a preferred embodiment of the present invention, the extracted feature are cepstral coefficients.

In box 153, a feature frame is formed which comprises the extracted features. The feature frame may include one or more digital words which represent the extracted features.

In box 154, the feature frame is distributed to a plurality of classifiers. Each of the classifiers implements a discriminant function which is tuned to indicate a different spoken sound. In response to receiving the feature frame, each classifier generates a classifier output signal which represents the result of computing the discriminant function.

In box 156, the identity of the spoken sound is determined by comparing the classifier output signals from the classifiers. In one embodiment of the present invention, the classifier output signal with the largest magnitude indicates the identity of the spoken sound.

In decision box 158, a check is made to determine if there is another spoken sound to be recognized from the continuously spoken speech. If there is another spoken sound, the method returns to box 150. If not, the method terminates.

Summary

There has been described herein a concept, as well as several embodiments including a preferred embodiment, of a both a method and system for recognizing spoken sounds from continuous speech.

Because the various embodiments of the present invention herein described utilize a plurality of classifiers they are insensitive to differences in speakers and not adversely affected by background noise.

It will also be appreciated that the various embodiments of the speech-recognition system as described herein do not require repetitive training; thus, the embodiments of the present invention require substantially

less training time and are significantly more accurate than known speech-recognition systems.

Furthermore, it will be apparent to those skilled in the art that the disclosed invention may be modified in numerous ways and may assume many embodiments other than the preferred form specifically set out and described above.

It will be understood that the concept of the present invention can vary in many ways. For example, it is a matter of design choice regarding such system structural elements as the number of classifiers, or the number of inputs to the selector. It is a matter of design choice whether the present invention is implemented in hardware or software. Such design choices greatly depend upon the integrated circuit technology, type of implementation (e.g. analog, digital, software, etc.), die sizes, pin-outs, and so on.

Accordingly, it is intended by the appended claims to cover all modifications of the invention which fall within the true spirit and scope of the invention.

What is claimed is:

CLAIMS

1. A system for recognizing spoken sounds from continuously spoken speech, the system comprising:

a plurality of classifiers, each of the classifiers receiving a feature frame and implementing a discriminant function to generate an output in response to the feature frame, wherein the feature frame is derived from the continuously spoken speech; and

a selector, responsive to the output of each of the classifiers, the selector identifying the spoken sound corresponding to an interval of time by comparing the classifier output signals from each of the plurality of classifiers.

2. The system recited in claim 1 further comprising:

a converter which receives the continuously spoken speech and generates a speech signal from the continuously spoken speech;

a feature extractor which is responsive to the speech signal and produces as output the feature frame by extracting a plurality of features corresponding to the continuously spoken speech at the interval of time; and

a means for distributing the feature frame to the plurality of classifiers.

3. The system of claim 1 wherein the plurality of features is selected from the group consisting of cepstral coefficients, predictive coefficients, and Fourier coefficients.

4. The system of claim 1 wherein each of the plurality classifiers implements a discriminant function which is tuned to indicate a different spoken sound.

5. The system of claim 1 wherein the discriminant function is based on a polynomial expansion.

6. The system of claim 5 wherein the polynomial expansion has the form:

$$y = \sum_{i=1}^{m} w_{i-1}\, x_1^{g_{1i}}\, x_2^{g_{2i}} \ldots x_n^{g_{ni}}$$

wherein y represents a dependent variable;

wherein i, m, and n are integers;

wherein w_{i-1} represents the coefficient for the ith term;

wherein $x_1, x_2, \ldots, x_n$ represent independent variables; and

wherein $g_{1i}, \ldots, g_{ni}$ represent the exponents for the ith term in the expansion which are applied to the independent variables.

7. The system of claim 1 wherein the feature frame comprises at least one digital word, the at least one digital word representing at least one of the plurality of features.

8. The system of claim 1 wherein the spoken sound is selected from the group consisting of word, character, and phoneme.

9. The system of claim 1 wherein the selector identifies a spoken sound corresponding to a sequence of intervals of time.

10. The system of claim 1 wherein the selector determines the classifier output signal with the largest magnitude.

FIG.1

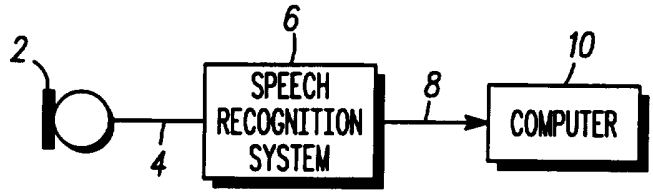


FIG.2

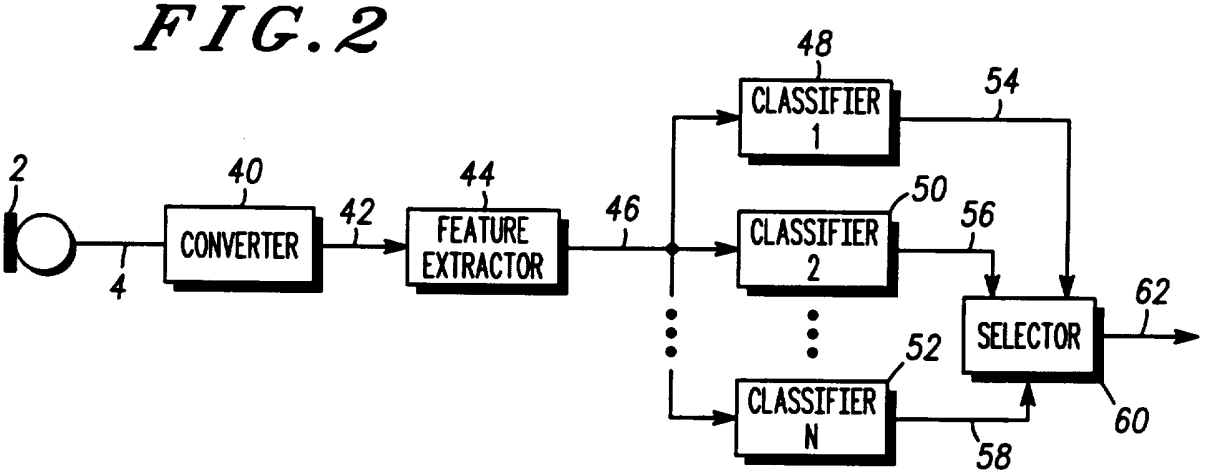
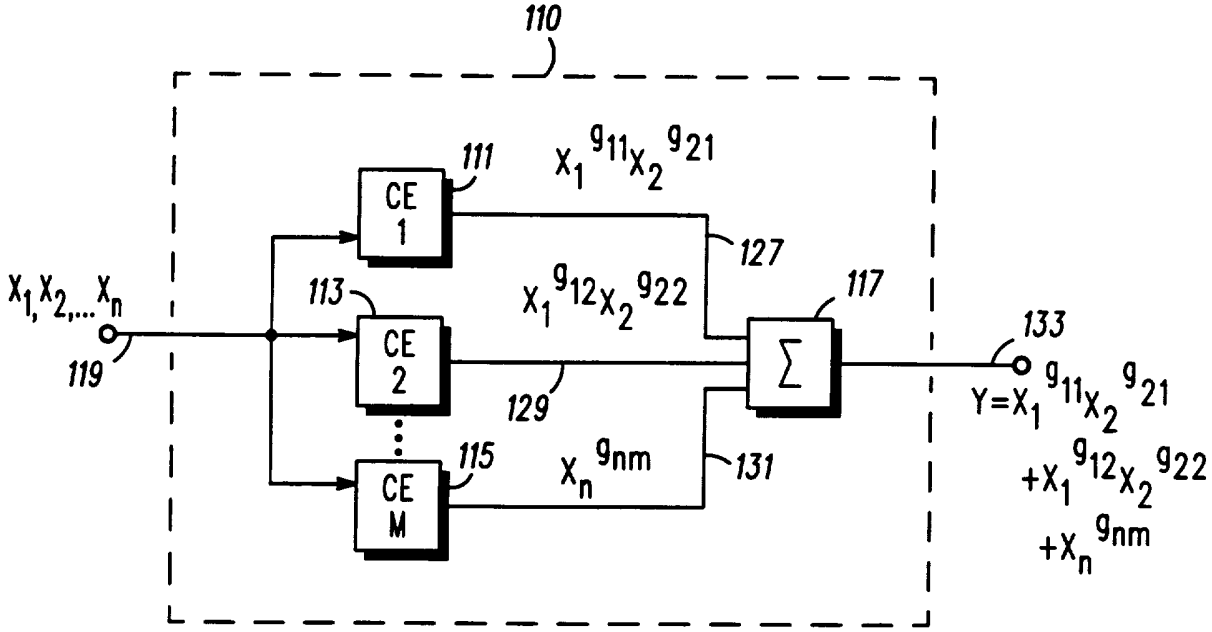


FIG.3



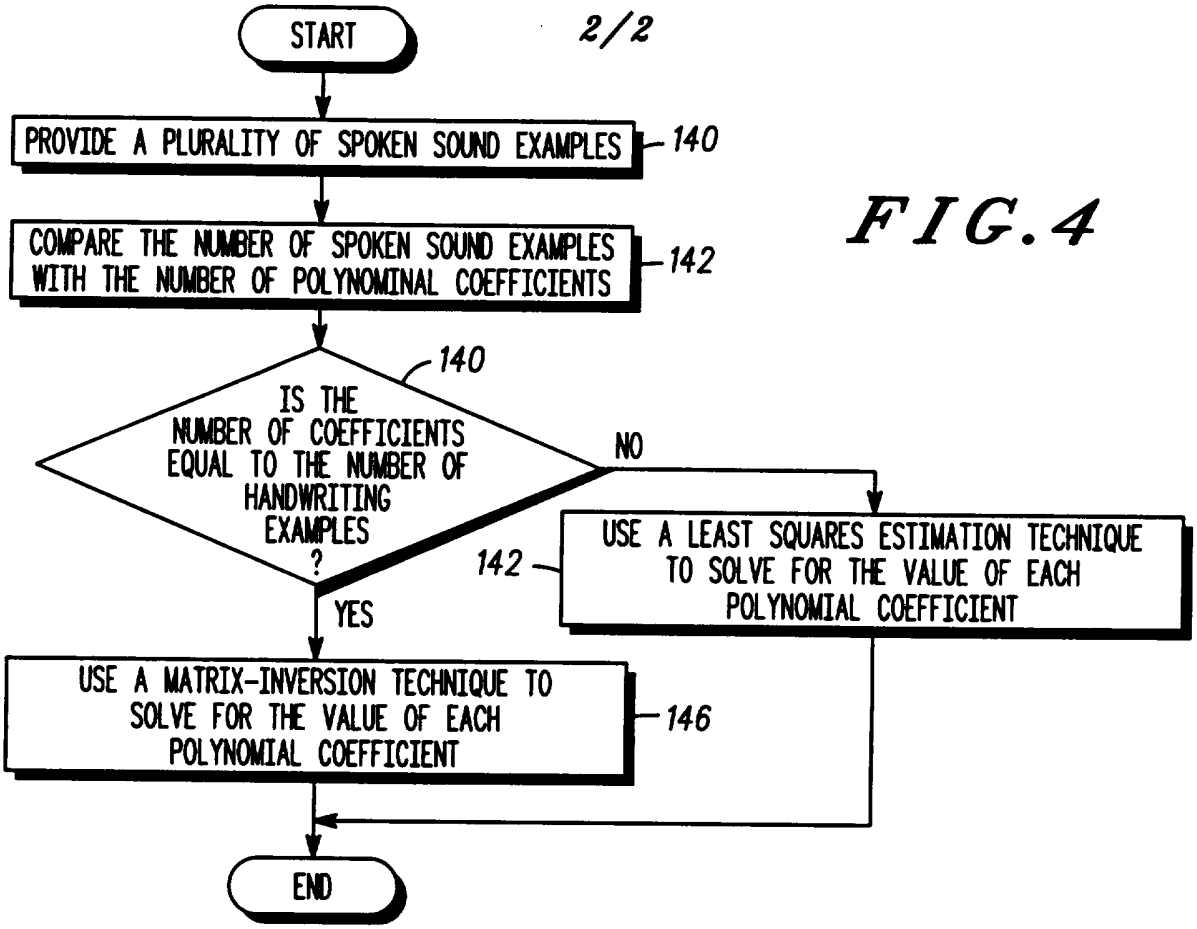

START
PROVIDE A PLURALITY OF SPOKEN SOUND EXAMPLES
140
COMPARE THE NUMBER OF SPOKEN SOUND EXAMPLES WITH THE NUMBER OF POLYNOMINAL COEFFICIENTS
142
140
IS THE NUMBER OF COEFFICIENTS EQUAL TO THE NUMBER OF HANDWRITING EXAMPLES ?
NO
142
USE A LEAST SQUARES ESTIMATION TECHNIQUE TO SOLVE FOR THE VALUE OF EACH POLYNOMIAL COEFFICIENT
YES
USE A MATRIX-INVERSION TECHNIQUE TO SOLVE FOR THE VALUE OF EACH POLYNOMIAL COEFFICIENT
146
END

FIG. 4

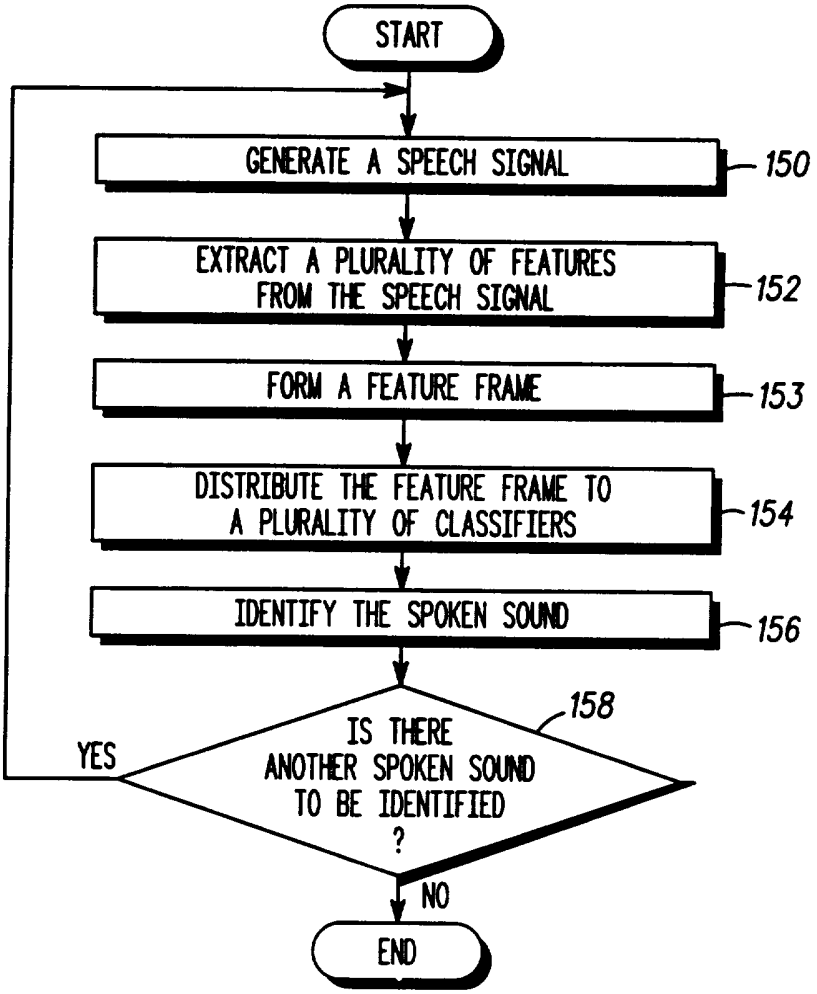

START
GENERATE A SPEECH SIGNAL
150
EXTRACT A PLURALITY OF FEATURES FROM THE SPEECH SIGNAL
152
FORM A FEATURE FRAME
153
DISTRIBUTE THE FEATURE FRAME TO A PLURALITY OF CLASSIFIERS
154
IDENTIFY THE SPOKEN SOUND
156
158
IS THERE ANOTHER SPOKEN SOUND TO BE IDENTIFIED ?
YES
NO
END

FIG. 5

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US95/08869

A. CLASSIFICATION OF SUBJECT MATTER

IPC(6) :G10L 5/00
US CL : 395/2.40, 2.45, 2.63, 2.64
According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

U.S. : 395/2.40, 2.45, 2.63, 2.64, 2.48, 2.54

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

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

APS, IEE/IEEE CD ROM

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	US, A, 5,299,284 (ROY) 29 March 1994, col. 4, lines 9-15; col. 3, lines 19-35	1-10
Y	US, A, 4,852,172 (TAGUCHI) 25 July 1989, Fig. 2	1-10

☐ Further documents are listed in the continuation of Box C. ☐ See patent family annex.

* Special categories of cited documents:	"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
"A" document defining the general state of the art which is not considered to be part of particular relevance	"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"E" earlier document published on or after the international filing date	"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
"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)	"&" document member of the same patent family
"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	

Date of the actual completion of the international search 16 SEPTEMBER 1995	Date of mailing of the international search report 03 NOV 1995
Name and mailing address of the ISA/US Commissioner of Patents and Trademarks Box PCT Washington, D.C. 20231 Facsimile No. (703) 305-3230	Authorized officer THOMAS ONKA Telephone No. (703) 305-9600

Form PCT/ISA/210 (second sheet)(July 1992)★