Description

Technical Field

This invention relates to speech recognition systems and more particularly to the types of such systems which are speaker independent.

Background of the Invention

It has been generally recognized that speech recognition is the most difficult of the three generic kinds of problems in the speech processing area. Speech coding of course is by now the most commercial type of speech processing equipment; and there is also some commerical equipment for speech synthesis.

Nevertheless, the speech recognition problem has been intractable to a large degree. Most recognition systems have been restricted in their ability to recognize speech from different speakers or to recognize more than an extremely limited vocabulary in an extremely focused or task-directed environment.

It has also been widely recognized that it would be desirable to have a speech recognition system that was capable of continuous speech recognition.

In recent years several word-based continuous speech recognition systems have been built. For example, one such system that has been built is that described by S. E. Levinson and L. R. Rabiner "A Task-Oriented Conversational Mode Speech Understanding System", *Speech and Speaker Recognition*, M. R. Schroeder, Ed., Kargar, Based Switzerland, pp. 149-96, 1985. That system and other similar systems recently developed are word-based in the first instance. While these systems have all been successful in their ability to accurately recognize speech in certain restricted ways, there is reason to believe that their use of words as the fundamental acoustic patterns precludes the possibility of relaxing the constraints under which they presently operate so that they can accept fluent discourse of many speakers over large vocabularies.

An often suggested alternative to the word-based approach is the so called acoustic/phonetic method in which a relatively few short-duration phonetic units, out of which all words can be constructed, are defined by their measurable acoustic properties. Generally speaking, speech recognition based on this method should occur in three stages. First, the speech signal should be segmented into its constituent phonetic units which are then classified on the basis of their measured acoustic features. Second, the phonetic units should then be combined to form words on some basis, using in part a lexicon which describes all vocabulary words in terms of the selected phonetic units. Third, the words should be combined to form sentences in accordance with some specification of grammar.

Several quite different embodiments of this basic methodology can be found in the literature if one is diligent to search for the less obvious aspects of some of the above components. See for example, the article by W.A. Woods, "Motivation and Overview of SPEECHLIS: An Experimental Prototype for Speech Understanding Research", *IEEE Transactions on Acoustics, Speech and Signal Processing*, Vol. ASSP-23, No. 1, February 1975, pp. 2-10. All such systems are very complex systems in which the basic components as outlined above are always present yet sometimes in disguised form. The fact that all these elements are not fully capable with respect to all syntax and with respect to all vocabulary means that they are very restricted in their use (e.g. task-oriented, as, for instance, retrieval of lunar rock sample information).

In "Explicit modelling of state occupancy in hidden Markov models for automatic speech recognition", ICASSP 85, pages 1.2.1 to 1.2.4, M.J. Russel and R.K. Moore disclose the use of Markov models with explicit state duration probability density functions in the modelling of words in a vocabulary for the purposes of speech recognition.

Accordingly, it is an object of this invention to provide an organization of such systems which is powerful in its ability to encompass all grammar and vocabulary and all speakers and is methodically organised so that it can be readily expanded.

Summary of the Invention

According to the present invention there is provided a speech recognition method as set out in claim 1.

My invention provides a powerful organization for further development, elaboration and expansion because, unlike the speech recognition systems described above, this one clearly and directly follows the outline of the acoustic/phonetic method as given above and is more capable of being implemented in the context of the most complicated applications. The directness and simplicity of the implementation derives

from the use of the continuously-variable-duration hidden Markov model as a model of the acoustic/phonetic structure of the language, instead of the less comprehensive set of phonological rules used in the prior art. The term "Markov" here refers to the probabilistic and not directly observable events which underlie the speech signal.

The continuously-variable-duration hidden Markov model employed in this invention represents the acoustic/phonetic structure of speech, rather than representing words directly. The phonetic units are identified with the unobservable states of the underlying Markov chain. The other parameters of the system are organized so that a dynamic programming algorithm can be used to obtain a phonetic transcription of an utterance.

In a preferred embodiment of my invention the lexical access phase of recognition is based on a pronouncing dictionary (above, termed a lexicon) in which each entry is orthographically represented in terms of phonetic symbols representing the states of the hidden Markov chain. Certain other features of the preferred embodiment will become apparent hereinafter.

The recognition process is then readily completed by prior art types of sentence recognition schemes by recovering from the word lattice the grammatically well-formed sentence of maximum likelihood conditioned on the measurements. It is expected that this final part of the recognition process will work more efficiently because of the candidate arrays and likelihood signals provided by the present invention.

Brief Description of the Drawing

FIG. 1 shows schematically the novel organization of the new acoustic/phonetic speech recognition system;

FIG. 2 shows schematically the prior art, of which the above-described Woods' article is representative;

FIG. 3 is a flow diagram for an optimization algorithm for detecting phonetic segments and durations, as required in the embodiment of FIG. 1;

FIG. 4 is a continuation of the flowchart of FIG. 3;

FIG. 5 is a block diagram representing some of the operations performed in the initial lexical access procedure for detecting words;

FIG. 6 is a diagram that illustrates some concepts employed in the flow chart of FIG. 7 as will be described below;

FIG. 7 is a flow chart relating to the alignment step;

FIG. 8 shows the concept of using pairs of adjacent phonetic segments in the alignment step; and

FIG. 9 is a flowchart of the implementation of a simple acoustic/phonetic model for the digits 0-9 as would be employed as a part of the lexical access procedures disclosed herein.

General Description

In the embodiment of FIG. 1 the input speech stream is illustratively derived from a human speaker; but the use of a speech stream from another machine is not precluded. The speech stream flows into equipment of known type shown as block 11 for linear predictive coding analysis (usually referred to as LPC analysis). The fundamental organization and composition of the equipment for the apparatus of block 11 is disclosed in some detail by F. Itakura in the article, "Minimum Prediction Residual Principle Applied to Speech Recognition", *IEEE Transactions on Acoustic Speech and Signal Processing*, Vol. ASSP-23, No. 1, February 1975. Nevertheless, with modern large scale integrated electronics several other implementations are possible as, for example, implied by FIG. 2 of my above-cited article with L. R. Rabiner.

The resulting sets of speech feature signals are then processed by a combination of techniques in the system portion 12 not only to detect phonetic segments with some degree of likelihood, but also to provide some valuable durational information with the arrays of segments. The algorithm for processing in apparatus 12 accesses a stored model 13 which is a so-called continuously-variable-duration hidden Markov model, as described in my article "Continuously Variable Duration Hidden Markov Models for Automatic Speech Recognition", *Computer Speech and Language*, Vol 1, pp. 29-44, Academic Press Inc., (London) Limited (1986). In addition, there is provided in portion 12 a searching and sorting technique which will be more fully described hereinafter and advantageously exploits complementary phonetic information. Next, with the candidate phonetic segments and durations, the system portion 14 accesses the lexicon 15 (an extensive memory electronically stored in a computer), preferably by pairs of phonetic segments, to obtain arrays of candidate words.

Then, an alignment procedure is carried out in system portion 16 in response to the acoustic features from the LPC analysis in block 11 to compute the likelihood signals for the indicated candidate words.

The preferred implementation of apparatus 12-16 is advantageously upon a digital computer, for example, of the type Alliant FX-8, manufactured by the Alliant Computer Corporation, Acton, Ohio. The model 13, the lexicon 15, and the sentence recognition apparatus 17 all use very substantial amounts of previously-determined information stored in memories therein.

The output of the alignment apparatus 16 is then fed into a prior art sentence recognition apparatus 17 of any of several types which resolves the remaining ambiguities among word candidates and selects the best sentence.

A basic understanding of the invention may be gained by considering the following points. One fundamental underpinning of the acoustic/phonetic embodiment proposed here is the assumption that speech is a code based on a small number of symbols. While the identity of these symbols is unknown, the physical manifestations of them can be observed directly by making measurements on the acoustic signal. Under this assumption and in light of experiments by many previous investigators, it is quite natural to model such a phenomenon as a so-called hidden Markov chain. Since the symbols of the code are assumed discrete, unknown and not directly observable, it is reasonable to identify them with the states of an underlying, unobservable probabilistic chain of events which chain of events is called commonly a Markov chain. On the other hand, the acoustic signal is accessible and its spectral and temporal features are presumed to be correlated with the symbols of the code. It is thus appropriate to represent the signal by a set of statistical distributions of acoustic observations and to associate one distribution with each state of the Markov chain. In order to account for temporal encoding of information in the signal we should also include a mechanism for characterizing the durations of the states probabilistically.

Detailed Description

It appears, then, that the continuously-variable-duration hidden Markov model rather neatly captures certain essential aspects of the speech signal. The phonetic units (i.e., code symbols) are represented by the hidden states

$$\{q_i\}_{i=1}^{n} .$$

The state transition probabilities, a_{ij}, of the hidden Markov chain account for the phonotactic structure of speech (i.e. the code itself). The acoustic correlates of the symbols are the time-indexed observations $O_t \in R^d$. The spectral information is contained in the set of observation densities,

$$\{b_j(O_t)\}_{j=1}^{n},$$

and the temporal information resides in the set of durational densities

$$\{d_j(t)\}_{j=1}^{n}.$$

The d-dimensional observation densities should be Gaussian and the durational densities, gamma distributed. Under these assumptions the parameters of an n-state model are the state transition probabilities, a_{ij}, $1 \le i,j \le n$; the observation means, μ_{jk}, $1 \le j \le n$, $1 \le k \le d$, the observation covariances, u_{jkl}, $1 \le j \le n$, $1 \le k,l, \le d$; and the durational shape and location parameters, ν_j and η_j $1 \le j \le n$. For this parametrization, methods are known for computing an optimal set of values. One such method will be referred to hereinafter.

In some speech recognition systems based on hidden Markov models, each vocabulary word is represented by a different model whose likelihood function can be evaluated for a given input and used in a Bayesian classification scheme. In my system, however, the state sequence plays the dominant role since it yields a phonetic transcription. The likelihood function is also used but for a less critical purpose, as we shall later see.

In the present acoustic/phonetic system the main issue, then, is how to determine the state sequence corresponding to a given speech signal. This is accomplished preferably by a dynamic programming algorithm that accounts for durational densities.

Let $\alpha_t(j)$ be the likelihood of the state sequence ending in state j which maximizes the joint likelihood of state and observation sequences up to and including time t. This quantity can be calculated recursively from

$$\alpha_t(j) = \max_{1\le i\le n}\left\{\max_{\tau\le t}\left\{\alpha_{t-\tau}(i)a_{ij}d_j(\tau)\prod_{\theta=1}^{\tau} b_j(O_{t-\tau+\theta})\right\}\right\}. \quad (1)$$

for $1 \le j \le n$ and $1 \le t \le T$
(The terms in equation 1 being defined in Appendix A).
If we simultaneously set

$$\beta_t(j) = (i,\tau) = \underset{i,\tau}{\text{argmax}}\ \{\alpha_t(j)\}, \quad (2)$$

then we can trace back from the final state

$$j_T = \underset{j}{\text{argmax}}\ \{\alpha_T(j)\} \quad (3)$$

to reconstruct the desired state sequence, thereby obtaining the phonetic transcription of the input.

The task of interpreting the phonetic transcription obtained from equations 1-3 as a sequence of words is not unlike that of reading a text corrupted by random insertions, deletions and substitutions of letters and stripped of spaces between words. The method by which we propose to render the interpretation is based on the psycholinguistic notion of lexical access.

The basic premise of the concept of lexical access is that words are stored in a dictionary where they are represented in some phonetic orthography rather than in terms of acoustic features. In the cognitive process of understanding speech, words are retrieved from the dictionary on the basis of partial phonetic information. The phonetic to lexical mapping is thus inherently many-to-one resulting in two kinds of ambiguities. First, phonetic sequences may be common to several words. Second, words may be retrieved on the basis of several different phonetic sequences. The first form of ambiguity makes the recognition problem more difficult and must be resolved at a latter stage of processing. The second kind, however, is beneficial. Since the phonetic transcription is corrupted, we need a multiplicity of cues for a given word. The particular lexical access procedure used in our invention advantageously provides a method for exploiting complementary phonetic information.

The particular form of the technique used here is as follows. Suppose that the phonetic transcription resulting from the evaluation of equations 1-3 is the state sequence

$$q = q_1, q_2 \cdots q_t, q_{t+1}, \ldots, q_{t+p}, \ldots, q_N. \quad (4)$$

where q_i is the number of the i^{th} state of the sequence for $1 \le i \le N$. We consider all $p + 1$ tuples of the form

$$q = q_t, q_{t+1}, \ldots, q_{t+p} \quad (5)$$

for $1 \le t \le N - p$ and fixed p. Each such q is taken as the argument of a hashing function,

$$f(q) = \left[\sum_{i=0}^{p} q_{t+i}\right] \text{mod} P = a \quad (6)$$

whose value, a, is an address in the lexicon. The prime number, P, is chosen to be greater than the number of entries in the lexicon which is arranged so that the a^{th} entry has an orthographic representation containing q as a subsequence. The lexicon is also constructed so that the a^{th} entry has a link to another word containing q. The last word in the chain has a null link. In general, each lexical item contains more than one $p + 1$ tuple and thus must have a separate link for each one. The links can all be generated

automatically from the lexical orthographic representations and need not be constructed manually.

Further details of the mathematics and computer implementation of such a technique in general can be found in the book by D. E. Knuth, *The Art of Computer Programming*, Addison-Wesley, Vol. 1, pp. 230-231, and Vol. 3, pp. 506-542, particularly those portions relating to "searching and sorting".

The lexical access procedure described above has the effect of giving coarse estimates of the locations of words, e.g., their beginnings and ends, in the speech stream. As noted, these locations are likely to be inconsistent. In order to resolve the ambiguity, we need to extract more information from the acoustic measurements and the phonetic transcription and represent that information in such a way that an optimal interpretation of the signal can be obtained. In particular, we wish to calculate the likelihood, $L(\nu|\tau_0, \tau_m)$, that the word ν appears in the time interval (τ_0, τ_m). Once computed for all words found by the access procedure, this information can usefully be represented by a graph called a word lattice.

Let us first consider a theoretically optimal approach to the problem. Suppose that the accessed word, ν, has orthography

$$\hat{q} = \hat{q}_1, \hat{q}_2, \ldots, \hat{q}_m$$

in which the subsequence matching the phonetic transcription

$$q = \hat{q}_l, \hat{q}_{l+1}, \ldots, \hat{q}_{l+p}.$$

is

Suppose, further, that we know that, at the very earliest, ν began at time τ_0 and ended no later than τ_m. Then $\hat{q}$ must have occurred somewhere in that interval with state $\hat{q}_1$ ending at some time $\hat{\tau}_1 \in (\tau_0, \tau_m)$ and of course, $\hat{\tau}_{i+1} > \hat{\tau}_i$ for $1 \le i \le m$. A reasonable definition of $L(\nu|\tau_0, \tau_m)$, then, is the maximum value of the joint likelihood of

$$O_{\hat{\tau}_0}, \ldots, O_{\hat{\tau}_m}$$

and $\hat{q}_1, \hat{q}_2, \ldots, \hat{q}_m$ over all choices satisfying the constraints of the state transition times $\hat{\tau}_0, \hat{\tau}_1, \ldots, \hat{\tau}_m$. That is,

$$L(v \mid \tau_0, \tau_m) = \max_{T} \left\{ L \left[O_{\hat{\tau}0} \cdots O_{\hat{\tau}_m}, \hat{q} \mid \tau_0, \tau_m \right] \right\} \quad (7)$$

where T is the set of all possible state transition times, namely

$$T = \{\hat{\tau}_0, \hat{\tau}_1, \ldots, \hat{\tau}_m \mid \tau_0 < \hat{\tau}_0 < \hat{\tau}_1 < \cdots < \hat{\tau}_m \le \tau_m\}. \quad (8)$$

As defined in (7) $L(\nu|\tau_0, \tau_m)$ can be computed from

$$\log[L(v \mid \tau_0, \tau_m)] = \max_{T} \left\{ \sum_{i=1}^{m} \left[\log a_{\hat{q}i-1\hat{q}i} + \sum_{t=\hat{\tau}_{i-1}}^{\hat{\tau}_i} \log b_{\hat{q}_i}(O_\tau) + \log d_{\hat{q}i}(\tau_i - \hat{\tau}_{i-1}) \right] \right\} \quad (9)$$

Although (9) is correct, it is practically useless since the cardinality of T is, in general, very large. There is, however, an efficient way to compute $L(\nu|\tau_0, \tau_m)$. Let $\gamma_t(\hat{q}_1)$ be the maximum joint log-likelihood of $O_{\tau 0} \cdots O_t$ and $\hat{q}_1, \hat{q}_2, \ldots, \hat{q}_1$. Then $\gamma_t(\hat{q}_1)$ can be calculated recursively for $\tau_0 < t \le \tau_m$ and $1 \le i \le m$ from

$$\gamma_t(\hat{q}_i) = \max_{\tau<a} \left\{ \gamma_{t-\tau}(\hat{q}_{i-1}) + \log a_{\hat{q}i-1\hat{q}_i} + \sum_{\theta=1}^{\tau} \log b_{\hat{q}i}\,(O_{t-\tau+\theta}) + \log d_{\hat{q}i}(\tau) \right\}. \qquad (10)$$

Finally,

$$\log[L(v \mid \tau_0, \tau_m)] = \gamma_{\tau_m}\,(\hat{q}_m) \qquad (11)$$

The lexical access and alignment procedure outlined above will generate many lexical hypotheses some of which will be inconsistent. We would like to retain all such hypotheses in a convenient form so that we can subsequently resolve the ambiguities and thereby recover the best sentence.

The data structure which meets the requirements is the word lattice. The horizontal direction in the lattice represents time and alternative lexical hypotheses are arranged vertically so that inconsistent hypotheses appear as sets of blocks having the same horizontal but different vertical positions. Such a data structure is conveniently stored in the computer as a directed graph whose vertices are word boundaries and whose edges are words weighted by their respective log likelihoods always directed forward in time. Let us denote the edge representing word ν during the interval from τ_i to τ_j having likelihood $L(\nu|\tau_i, \tau_j)$ by $(\tau_i, \nu, \tau_j, L_\nu)$. Of course, in general, there will be gaps and overlaps in the lattice which would render the corresponding graph disconnected. This problem is eliminated by identifying two distinct boundaries whenever their times of occurrence are within an appropriate, empirically chosen neighborhood of each other.

Now let us turn to the flow diagram of FIGS. 3 and 4. The first step in the operation of the central processor of the computer is to reset all of the central processing registers to the appropriate starting condition. The input data from apparatus 11 will come in as an array of quantities to be stored in an input register and we have designated this array as array b_j, as indicated in the flow diagram.

All processing through block 35 sets initial memory values. Blocks 36 and 37 increment state and time indices. Blocks 38-39 compute the sum of log probabilities of observations.

Blocks 40-44 compute the maximum likelihood of the first t observations originating in the initial state.

Blocks 45-54 compute the maximum probability of the first t observations originating from any state sequence.

Blocks 55-59 determine the most likely state from which the last observation was generated, and the likelihood of the observation sequence.

Block 30-end reconstructs the optimal state sequence.

All of the needed parameters for the first array of input data to procedures described above have now been calculated and the process then proceeds to work on the next array of input data. Each set of data resulting from the iterative cycling through flow diagram of FIGS. 3 and 4 can then be stored in an appropriate position in a memory for use in the final calculations of equations 1,2 and 3 above so that finally the phonetic transcription of the input is reconstructed from the desired state sequences of each cycle.

As mentioned above, the task of interpreting the resulting phonetic transcription is like reading a text in which the words are run together and there are random insertions, deletions and substitutions.

One particular form of the technique used here is illustrated in FIGS. 5 and 6.

For the special case of p = 1, the hash function, equation (6), can be replaced with the HEAD table as shown in FIG. 6. For the significance and implementation of the HEAD table, see Knuth, Vol. 3 (cited above). If $q = q_t q_{t+1} = i,j$ then the entry row i and column j of the HEAD table, r, is the address of the lexical entry containing q. This table is also easily constructed automatically. The HEAD table method can be generalized to p = 2 or even more, but eventually it becomes inefficient.

It is not necessary to use the definition of equation 7 nor the recursion of equation 10 because of the existence of the following utterly simple and highly effective heuristic. Let us simply define $L(\nu|\tau_0, \tau_m)$ to be the joint likelihood of $O_\tau 0 \ldots O_{\tau m}$ and $\hat{q}$, where the state transition times are determined by the state durations computed from equation 2. Thus we can easily compute $\log[L(v|\tau_0\tau_m)]$ from equation (9) without the maximization over T using, instead, the previously computed segmentation of the phonetic transcription.

The flowchart for accomplishing the most essential calculations of equations 4-10 is shown in FIG. 7. This flowchart together with the alignment diagram of FIG. 8 illustrates how the various results of the previous steps, now stored in various large-scale memories in the computer, cooperate in the alignment step. The initial access is shown in block 73. The alignment as shown in FIG. 8 is represented by blocks 75-77. Block 8 depicts the computation of the word likelihood from the following

$$\log L(v \mid \underline{q},\underline{O}) = \sum_{l=1}^{L-1} \left[\log a_{\underline{q}_l \underline{q}_{l+1}} + \sum_{t=\tau_{l-1}}^{\tau_l} b_{\underline{q}_l t} + \hat{d}_{\underline{q}_l} \right] \qquad (12)$$

N.B. $\underline{q}_R\ \underline{q}_{R+1} = \hat{\underline{q}}_l\ \hat{\underline{q}}_{l+1}$ for some k,l

and its storage on the candidate list. Block 80 shows how the linked list is used to access the next candidate word. Block 81 is a test for the last link in the list. When this test succeeds, control is returned to the state counter and the next state pair is presented to the HEAD table. Block 83 shows the test for the last state pair. When this test succeeds, control passes to block 84 in which the best candidate is chosen and the process terminates.

The alignment process shows how an estimated state sequence can be matched to a dictionary entry. In general the match will not be perfect but there will be multiple state sequences which will partially match the estimated state sequence of the candidate word.

It should be noted in the passing that other alignment algorithms are possible such as those based on dynamic programming even though it is our objective to avoid dynamic time-warping if possible. These other algorithms are compatible with the feature of my invention which has introduced the concept of using the separate alignment step indicated in FIG. 1.

One final comment on the alignment procedure of FIG. 7, is that, in principle, all parrs of adjacent states in $\hat{q}$ should be used as keys in the alignment process.

The details of the sentence recognition apparatus 17 are not an essential part of my invention. It is suggested that an appropriate organization of the sentence recognition apparatus could be that disclosed in my U. S. Patent No. 4,277,644, issued July 7, 1981.

As an alternative to the above referenced technique, it will be immediately apparent to many workers in the art that various other sentence recognition techniques can be used.

For recognizing the digits, 0-9, a 22 state model can be constructed, as shown in FIG. 9. To simplify the structure of the underlying Markov chain three states are replicated to bring the total to 25. Vectors associated with each segment are collected and maximum likelihood estimators for means, covariances, durations and standard deviations of duration are computed. The values obtained for a segment are then used as the parameter values associated with the corresponding state.

The transition probabilities of the underlying Markov chain are estimated on the assumption that all digits are equally likely. Given that assumption, it is a simple matter to count the number of times each transition will occur from which event counts the transition probabilities are easily obtained.

FIG. 9 shows the kinds of phonetic elements which speech recognition workers have come to accept as being the most likely representation of the irreducible elements of the hidden process occurring in the vocal trade, or, at least, the most convenient representations of those elements.

It will be immediately apparent to many workers in the art of various modifications that can be made upon the above-described techniques depending on the level of generality and capability versus the level of complexity required in a system according to my present invention.

APPENDIX A

q_j = state j

O_t = feature vector at time t

α_{jt} = joint log likelihood of $O_1 O_2 \cdots O_t$ and $\hat{q}_1 \hat{q}_2 \cdots \hat{q}_k$

β_{jt} = pair of pointers to previous state and time

$S_j\tau = \sum_{\theta}^{\tau} 1 b_{jt-\tau+\theta}$

δ = maximum duration of a state

b_{jt} = log likelihood of O_t given q_j

d_{jt} = log likelihood of duration t given q_j

a_{ij} = probability of transition from q_1 to q_j

π_j = probability of q_j at t=0

p = joint log likelihood of $O_1 O_2 \cdots O_t$ and $\hat{q}_1 \hat{q}_2 \ldots \hat{q}_l$

$\hat{q}_j$ = j^{th} state in estimated state sequence

$\hat{d}_j$ = duration of $\hat{q}_j$

v = candidate word

W_r = r^{th} entry in the lexicon

$\hat{W}$ = is the best candidate word

C_n = the n^{th} candidate word

Claims

1. A method for the recognition of speech, of the type including the steps of
storing signals representing a model of the language to be recognized, said model being of the

* C L M S *

APPENDIX A

q_j = state j

$\underline{O}_t$ = feature vector at time t

α_{jt} = joint log likelihood of $O_1 O_2 \cdots O_t$ and $\hat{q}_1 \hat{q}_2 \cdots \hat{q}_k$

β_{jt} = pair of pointers to previous state and time

$S_j\tau = \sum_{\theta=1}^{\tau} b_{jt-\tau+\theta}$

δt = maximum duration of a state

b_{jt} = log likelihood of $\underline{O}_t$ given q_j

d_{jt} = log likelihood of duration t given q_j

a_{ij} = probability of transition from q_1 to q_j

π_j = probability of q_j at $t=0$

p = joint log likelihood of $O_1 O_2 \cdots O_t$ and $\hat{q}_1 \hat{q}_2 \ldots \hat{q}_l$

$\hat{q}_j$ = j^{th} state in estimated state sequence

$\hat{d}_j$ = duration of $\hat{q}_j$

v = candidate word

W_r = r^{th} entry in the lexicon

$\hat{W}$ = is the best candidate word

C_n = the n^{th} candidate word

Claims

1. A method for the recognition of speech, of the type including the steps of
storing signals representing a model of the language to be recognized, said model being of the

state-transitional type, each state being uniquely identified with a phonetic unit,

each state having associated with it a portion of a transition matrix which describes which states can follow it and with what probability,

each state having associated with it an observational density function assigning to each set of speech feature signals that may be observed in fluent speech a likelihood of being observed in association with that state,

each state having associated with it a durational density function assigning to each duration it may have a likelihood of occurrence in fluent speech;

storing signals representing lexical candidates, said lexical candidates being assemblages of phonetic units of the language in association with the partial phonetic information of the type found in dictionaries;

sequentially converting successive time frame portions of an utterance occurring during corresponding successive time frames into signals representing respective sets of acoustic feature signals representative of the portions; and

accessing the stored model and stored lexical candidates to obtain signals which represent sequences of the phonetic units, including selecting the optimum ones of such sequences to produce a selection signal representing recognition of the utterance.

CHARACTERIZED IN THAT

the accessing step includes

assigning a phonetic unit signal and a phonetic duration signal from the stored model to one or more of said time frame portions of speech as represented by said one or more of said respective sets of acoustic feature signals;

maximizing independently of the stored lexical candidates the likelihoods of each phonetic unit and each phonetic duration jointly with the likelihood of observing said one or more of said respective sets of acoustic feature signals;

accessing the stored lexical candidates corresponding to portions of the proposed sequences of phonetic units to obtain proposed lexical candidates each containing each said portion;

whenever multiple proposed lexical candidates contain the same one said portion, aligning said multiple proposed lexical candidates with the one or more respective sets of acoustic feature signals from which said proposed sequences of phonetic units were obtained to evaluate said multiple proposed lexical candidates.

2. A method for the recognition of speech of the type claimed in claim 1,

said method being further characterized in that the model storing step includes storing an ergodic model in which any state can occur after any other state.

3. A method for the recognition of speech of the type claimed in claim 2,

said method being further characterized in that

the lexical candidate storing step comprises

storing words represented by a phonetic orthography which is characterized by partial phonetic information, such that words may be retrieved on the basis of the phonetic units which they contain, and

storing information linking a plurality of words containing like sequences of phonetic units, whereby for each sequence of phonetic units as many words as contain them are directly accessible.

4. The method for the recognition of speech of the type claimed in claim 1, said method being further characterized in that

said assigning and maximizing are performed recursively for all assignments and transitions over all of said time frames up to and including the present time frame; and

the results of said maximizing are retraced by stepping through the phonetic determinations in a strict order to produce a proposed phonetic sequence for accessing the lexical candidates.

5. A method for the recognition of speech of the type claimed in claim 1, said method being further characterized in that

the aligning step evaluates said multiple proposed lexical candidates to include only those suitable for subsequent unambiguous ranging by processing by techniques that relate to sentence structure and meaning.

* C L M S *

6. A method for the recognition of speech of the type claimed in claim 1 further characterized in that
the aligning step evaluates said multiple proposed lexical candidates to select only the best one, whereby a selection signal representing the utterance as a word is produced.

Patentansprüche

1. Verfahren zur Erkennung von Sprache mit den Schritten:
Speichern von Signalen, die ein Modell der zu erkennenden Sprache darstellen, wobei das Modell vom Zustandsübergangstyp ist und jeder Zustand eindeutig durch eine phonetische Einheit identifiziert wird,
jedem Zustand ein Teil einer Übergangsmatrix zugeordnet ist, die Zustände beschreibt, denen der Zustand folgen kann und mit welcher Wahrscheinlichkeit,
jedem Zustand eine Beobachtungsdichtefunktion zugeordnet ist, die jedem Satz von Sprachmerkmalssignalen, die sich in flüssiger Sprache beobachten lassen, eine Wahrscheinlichkeit für die Beobachtung in Verbindung mit diesem Zustand zuweist, und
jedem Zustand eine Dauerdichtefunktion zugeordnet ist, die jeder Dauer, die der Zustand haben kann, eine Wahrscheinlichkeit für das Auftreten in flüssiger Sprache zuweist,
Speichern von Signalen, die lexikalische Kandidaten darstellen, welche Ansammlungen phonetischer Einheiten der Sprache in Verbindung mit phonetischen Teilinformationen des Typs sind, der sich in Wörterbüchern findet,
sequentielles Umwandeln aufeinanderfolgender Zeitrahmenteile einer Äußerung, die in entsprechend aufeinanderfolgenden Zeitrahmen auftreten, in Signale, die entsprechende Sätze akustischer, die Teile darstellender Merkmalssignale darstellen, und
Zugreifen auf das gespeicherte Modell und die gespeicherten lexikalischen Kandidaten zur Gewinnung von Signalen, die Folgen der phonetischen Einheiten darstellen, einschließlich Auswählen der optimalen Mitglieder dieser Folgen zur Erzeugung eines Auswahlsignals, das die Erkennung der Äußerung darstellt,
dadurch gekennzeichnet, daß der Zugriffsschritt die Schritte enthält:
Zuordnung eines Phonetikeinheit-Signals und eines Phonetikdauer-Signals aus dem gespeicherten Modell zu einem oder mehreren der Zeitrahmenteile von Sprache, dargestellt durch einen oder mehrere der entsprechenden Sätze akustischer Merkmalssignale,
Maximieren unabhängig von den gespeicherten lexikalischen Kandidaten der Wahrscheinlichkeiten jeder phonetischen Einheit und jeder phonetischen Dauer zusammen mit der Wahrscheinlichkeit für die Beobachtung des einen oder mehrerer entsprechender Sätze akustischer Merkmalssignale,
Zugreifen auf die gespeicherten lexikalischen Kandidaten entsprechend Teilen der vorgeschlagenen Folgen phonetischer Einheiten zur Gewinnung vorgeschlagener lexikalischer Kandidaten, die je jeden Teil enthalten,
immer dann, wenn mehrere vorgeschlagene lexikalische Kandidaten den gleichen Teil enthalten, Zuweisen der mehreren vorgeschlagenen lexikalischen Kandidaten zu dem einen oder den mehreren entsprechenden Sätzen akustischer Merkmalssignale, aus denen die vorgeschlagenen Folgen phonetischer Einheiten gewonnen worden sind, um die mehreren vorgeschlagenen lexikalischen Kandidaten auszuwerten.

2. Verfahren zur Erkennung von Sprache nach Anspruch 1,
dadurch gekennzeichnet, daß der Modellspeicherschritt die Speicherung eines ergodischen Modells beinhaltet, bei dem jeder Zustand nach jedem anderen Zustand auftreten kann.

3. Verfahren zur Erkennung von Sprache nach Anspruch 2,
dadurch gekennzeichnet, daß der Schritt zur Speicherung lexikalischer Kandidaten die Schritte umfaßt:
Speichern von Wörtern, die durch eine phonetische Orthographie dargestellt werden, welche durch partielle phonetische Informationen gekennzeichnet ist, derart, daß Wörter auf der Grundlage der phonetischen Einheiten wiedergewonnen werden können, die sie enthalten, und
Speicherung von Informationen, die eine Vielzahl von gleiche Folgen phonetischer Einheiten enthaltender Wörter verketten, wodurch für jede Folge phonetischer Einheiten so viele Wörter direkt zugreifbar sind, wie sie enthält.

4. Verfahren zur Erkennung von Sprache nach Anspruch 1,
dadurch gekennzeichnet, daß die Zuordnung und Maximierung rekursiv für alle Zuordnungen und

* C L M S *

Übergänge über alle Zeitrahmen bis zum und einschließlich des augenblicklichen Zeitrahmens durchgeführt wird und
daß die Ergebnisse der Maximierung mittels eines Durchlaufens der phonetischen Bestimmungen in einer strengen Reihenfolge zurückverfolgt werden, um eine vorgeschlagene phonetische Folge für einen Zugriff zu den lexikalischen Kandidaten zu erzeugen.

5. Verfahren zur Erkennung von Sprache nach Anspruch 1,
dadurch gekennzeichnet, daß der Ausrichtungsschritt die Vielzahl von vorgeschlagenen lexikalischen Kandidaten so auswertet, daß sie nur diejenigen enthalten, die für eine nachfolgende unzweideutige Reihenfolge geeignet sind, und zwar durch eine Verarbeitung durch Verfahren, die sich auf die Satzstruktur und die Bedeutung beziehen.

6. Verfahren zur Erkennung von Sprache,
dadurch gekennzeichnet, daß der Ausrichtungsschritt die Vielzahl von vorgeschlagenen lexikalischen Kandidaten so auswertet, daß nur der beste Kandidat ausgewählt wird, wodurch ein Auswahlsignal erzeugt wird, das die Aussprache als ein Wort darstellt.

Revendications

1. Une méthode pour la reconnaissance de la parole, du type comprenant les étapes de
mémorisation de signaux représentant un modèle du langage à reconnaître, ledit modèle étant du type à état transitionnel, chaque état étant identifié de manière unique par une unité phonétique,
l'association à chaque état d'une partie d'une matrice de transition qui décrit les états pouvant le suivre et avec quelle probabilité,
l'association à chaque état d'une fonction de densité observationnelle assignant à chaque jeu de signaux de caractéristiques de parole pouvant être observé dans une conversation courante une vraisemblance d'être observé en association à cet état,
l'association à chaque état d'une fonction de densité de durée assignant à chaque durée qu'il pourrait avoir une vraisemblance d'occurrence dans la conversation courante;
la mémorisation de signaux représentant des candidats lexiques, lesdits candidats lexiques étant des assemblages d'unités phonétiques du langage en association avec les informations phonétiques partielles du type trouvé dans les dictionnaires;
la conversion séquentielle des parties de tranche de temps successives d'une émission d'un son se produisant durant les tranches de temps successives correspondantes en signaux représentant des jeux respectifs de signaux de caractéristiques acoustiques représentatifs des parties; et
l'accès au modèle mémorisé et aux candidats lexiques mémorisés pour obtenir des signaux qui représentent des séquences des unités phonétiques, comprenant la sélection des séquences optimales de ces séquences pour produire un signal de sélection représentant la reconnaissance de l'émission du son,
CARACTERISE EN CE QUE
l'étape d'accès comprend
l'assignation d'un signal d'unité phonétique et d'un signal de durée phonétique du modèle mémorisé à une ou plusieurs desdites parties de tranche de temps de parole telles qu'elles sont représentées par lesdits un ou plusieurs jeux respectifs de signaux de caractéristiques acoustiques;
la maximisation indépendamment des candidats lexiques mémorisés des vraisemblances de chaque unité phonétique et de chaque durée phonétique conjointement avec la vraisemblance d'observation desdits un ou plusieurs jeux respectifs de signaux de caractéristiques acoustiques;
l'accès aux candidats lexiques mémorisés correspondant aux parties des séquences proposées d'unités phonétiques pour obtenir les candidats lexiques proposés contenant chacun ladite partie;
chaque fois que les candidats lexiques multiples proposés contiennent la même dite partie, l'alignement desdits candidats lexiques multiples proposés avec le ou les jeux respectifs de signaux de caractéristiques acoustiques à partir desquels lesdites séquences proposées d'unités phonétiques ont été obtenues pour évaluer lesdits candidats lexiques multiples proposés.

2. Une méthode pour la reconnaissance de la parole du type revendiqué à la revendication 1,
ladite méthode étant encore caractérisée en ce que l'étape de mémorisation de modèle comprend la mémorisation d'un modèle ergodique dans lequel n'importe quel état peut se produire après n'importe quel autre état.

3. Une méthode pour la reconnaissance de la parole du type revendiqué à la revendication 2,
ladite méthode étant encore caractérisée en ce que
l'étape de mémorisation des candidats lexiques comprend
la mémorisation de mots représentés par une orthographe phonétique qui est caractérisée par des informations phonétiques partielles, de telle sorte que les mots peuvent être extraits d'après les unités phonétiques qu'ils contiennent, et
la mémorisation d'informations liant une pluralité de mots contenant des séquences semblables d'unités phonétiques, par laquelle pour chaque séquence d'unités phonétiques, le plus grand nombre de mots les contenant sont directement accessibles.

4. La méthode pour la reconnaissance de la parole du type revendiqué à la revendication 1, ladite méthode étant encore caractérisée en ce que
lesdites assignation et maximisation sont exécutées de manière récurrente pour toutes les assignations et transitions sur toutes lesdites tranches de temps jusqu'à et y compris la tranche de temps actuelle; et
les résultats de ladite maximisation sont retracés en examinant les déterminations phonétiques dans un ordre strict pour produire une séquence phonétique proposée pour accéder aux candidats lexiques.

5. La méthode pour la reconnaissance de la parole du type revendiqué à la revendication 1, ladite méthode étant encore caractérisée en ce que
l'étape d'alignement évalue lesdits candidats lexiques multiples proposés pour inclure seulement ceux convenant à une classification non ambiguë subséquente par des sous-techniques de traitement qui se rapportent à la structure et au sens de la phrase.

6. La méthode pour la reconnaissance de la parole du type revendiqué à la revendication 1 encore caractérisée en ce que
l'étape d'alignement évalue lesdits candidats lexiques multiples proposés pour ne sélectionner que le meilleur, de manière à produire un signal de sélection représentant l'émission d'un son comme un mot.