DRAGON DICTATE™-30K: NATURAL LANGUAGE SPEECH RECOGNITION WITH 30,000 WORDS

Janet M. Baker
Dragon Systems, Inc.
90 Bridge Street
Newton, MA 02158
U.S.A.
TEL: (617) 965-5200
FAX: (617) 527-0372

The DragonDictate™-30K automatic transcription system is a large vocabulary, discrete utterance, natural language speech recognition system. First-time users are not required to provide any initial training or enrollment speech data before using the system. Speech models for 16,000 frequent words/phrases of English are built-in. Powerful adaptation techniques dynamically build and refine up to 30,000 speech models on-line during system use. With its integration of a 80,000 word dictionary and easy introduction of new terminologies, DragonDictate™ offers a large, open vocabulary which can be used to generate free text immediately in any discipline, however specialized. Although slower than a skilled typist, interactive users of the DragonDictate™ system typically create printed text more quickly by speaking than they can generate handwritten text alone. Compatible with most popular and custom software for word-processing, databases, spreadsheets, etc., DragonDictate™-30K presently consists of 1 8-bit peripheral card and software, running near real-time on an MS-DOS 386-based (AT-bus), personal computer with 6 MB memory. Professionals, business executives, and other people with limited typing skills, are now independently creating their documents and reports simply by speaking them. This paper describes the features of DragonDictate™ from both technical and user perspectives.

SYSTEM DESIGN

The DragonDictate™ speech recognizer is based on a Multiple Knowledge Source (MKS) system design. This philosophy supports the integration of different types and levels of information into a consistent multi-disciplinary framework. Its stochastic modeling methodology allows for the rigorous evaluation of competing inter- and intra-word hypotheses. The principle of "delayed decisions" implicit in stochastic processing enables the maximum utilization of available information from these different knowledge sources in the resolution of ambiguity. Although this design methodology is generally applicable for a variety of speech recognition tasks, it is especially valuable for very large vocabulary recognition. No single knowledge source suffices to resolve with high reliability, the ambiguity inherent in simultaneously active vocabularies of thousands to tens of thousands of words explicitly trained to an individual speaker. Because the DragonDictate™-30K transcription system starts with 16,000 speaker-independent words simultaneously active, the a priori potential for ambiguity is very high. Nonetheless cooperative multiple knowledge sources, especially utilized in conjunction with interactive "automatic learning" or adaptation techniques, have proved highly effective even in this challenging scenario.

SYSTEM COMPONENTS

The MKS approach applied at the system level is also applied, where appropriate, to individual system components or functional units (e.g. signal processing, acoustic-phonetic models, language modeling, etc.). Given a highly modular structure, DragonDictate™ readily lends itself to experimentation with alternative versions within individual components to change, for example, the sampling rate, number and type of signal parameters, phone and word models, language model statistics, etc. Depending on the constraints of a specific system or application configuration, different choices for these may be preferred.

The analog signal from the microphone (e.g. noise-canceling, headset) is digitized and sampled (4000 Hz minimum). Signal parameters (e.g. peak amplitude) observed in the time domain are combined with frequency domain parameters to create a series of frames (20 msec. maximum), each containing a vector of these combined parameters. The frequency domain analysis is performed by an FFT analysis (8 band minimum) using a Hanning window. An utterance detector compares peak amplitude in a succession of frames against dynamically modified background noise estimations to determine the start time of a putative utterance, and to initiate host computer processing.
Incoming frames are matched against a series of frame-like nodes in the models for each active vocabulary word. Separate probability distributions are associated with each frame parameter. A Laplacian probability can then be computed on a parameter by parameter basis, and combined to yield a likelihood score of a match between that frame and any corresponding word node model. Initial word models can be constructed from one or more actual utterances (from the same or other speaker(s)), or hypothesized on the basis of concatenated acoustic-phonetic elements, given one or more base pronunciations comprised of collections of these elements. Successive utterance instances can easily be used to refine and improve word models automatically.

For each active node, a score is produced corresponding to the sum of the probabilities of all dynamic programming paths which result in the alignment of the current frame with this node. Successive frames are aligned with successive word model nodes, and for each incoming frame, active node scores are recalculated to sum the probabilities of that frame's remaining in the same node, a "self-loop", as compared to transitioning to a successive node. Durational penalties are applied to the self-loop probability if a given frame remains aligned with a particular node for longer than that node's expected duration. Similarly, penalties are assessed for unexpectedly rapid transitions to successive nodes.

For each active vocabulary word considered, its acoustic likelihood match is combined with its language model score. Recognition of utterances and their respective rank-orderings are determined by these combined scores. The language model used is also statistical in nature. Large text databases (e.g. UPI newswire data) have been extensively processed and analyzed to extract n-gram statistics as well as application specific linguistic information (reflecting more specific semantic and syntactic knowledge sources), where appropriate (e.g. for specialized medical domains). As additional data is made available to DragonDictate™, the language model can be improved and refined automatically. The system easily accommodates to large variations in the strength of the language model, as can be observed, worst case, in the recognition of random word sequences, or even a foreign language where there exist no appropriate acoustic word models or language statistics. Given its on-line adaptation capabilities, the DragonDictate™ system, with consistent use, will automatically modify its internal models and introduce new models (both acoustic and linguistic models) to recognize these new speech genres with high accuracy.

Important computational savings are achieved using a "rapid match" to determine which words, out of all the words possible, are most likely to match a spoken utterance. A more expensive detailed match can then be performed on this rapid match vocabulary subset. This rapid scoring averages acoustic parameters from multiple frames from a spoken utterance and compares these averaged parameters with a corresponding set of averaged parameters for all words initially active. Partial language model scores can also be incorporated into the rapid match computation. During the detailed match, the rapid match vocabulary words are scored in parallel to permit additional pruning when a given word candidate's score deviates from the best scoring word by more than a specified amount.

OPEN VOCABULARY WITH SEAMLESS BOUNDARIES

As with a completely speaker-independent system, a first-time user may start speaking to the DragonDictate™ system without initially providing any speech data for training or enrollment. Through the dynamic adaptation occurring automatically during interactive speech recognition, the system quickly builds and refines speaker-dependent models to achieve optimal performance. DragonDictate™-30K presently starts with 16,000 simultaneously active speaker-independent word patterns. The built-in Random House® Unabridged Dictionary has 80,000 words available in a well integrated user interface. New words are easily added to the active vocabulary, whether or not they are in the dictionary. In the event that new words are introduced beyond the maximum active vocabulary capacity, i.e. 30,000 words for the DragonDictate™-30K, these new words will automatically replace the least recently used words in the active vocabulary.

The "open vocabulary" construct introduced by DragonDictate™ refers to the capability of allowing out-of-vocabulary words and dynamically adding them to the active vocabulary. "Seamless boundaries" between vocabularies refers to the fact that a user doesn't need to know which words are in the active vocabulary, or trained, or in the dictionary. The vocabulary size characteristic for DragonDictate™ featuring an open vocabulary with seamless boundaries, will vary for users over time, along several kinds of vocabulary dimensions. For a given user at a single point in time, the system may, for example, have 8000 words in its speaker-dependent vocabulary, 21,000 words in its active vocabulary, 64,000 words in its model vocabulary, 80,000 words in its spelling vocabulary, and an unbounded number of words in its total vocabulary.

Extensive analysis of the Random House® Unabridged Dictionary has provided over 12,000 distinctive phone contexts (typically though not exclusively, associated with phone triplets). These diverse phone contexts are reflected in the model vocabulary. Disambiguating acoustically similar words often depends on accurately resolving such allophonic phenomena.

SYSTEM SPECIFICATIONS

The DragonDictate™-30K system consists of a single full length 8-bit peripheral card (incorporating a TMS 32010 DSP) and recognition / interface / dictionary software (plus microphone, documentation, etc.). This system presently requires an MS-DOS 80386 personal computer with an AT bus and 6 MB RAM (divided into expanded and extended memory segments). The DragonDictate™ speech driver runs in protected mode. Separate user files for multiple speakers can be stored on the same system hard disk.

USER INTERFACE

Some of the most important advances of DragonDictate™ over previous systems, are not directly related to its speech recognition technology. The DragonDictate™ system is simply much easier to use across a variety of both horizontal and vertical applications. Eliminating initial user training / enrollment procedures allows new users immediate system access. To use DragonDictate™ you can simply 1) load your DragonDictate™ software, 2) load your word processor, and 3) start dictating! As you speak each word, a menu appears with the best scoring word appearing at the top and close choices displayed below. If the top choice is correct, simply continue on to speak the next word. If the correct word is displayed, but not as the top choice, you may select it with a corresponding function key (i.e. choose the 3rd word by hitting the F3 key) or verbally (i.e. by speaking "pick 3").
If the correct word is not displayed, then you start typing, or verbally spelling, the word. Often the correct word will then be displayed after the first few letters are entered, and you then select the correct word from the menu as described above. The most that is required, even to enter a new vocabulary word, is simply to say it and spell it once. Successive menus display the best word choices derived from all levels of the available vocabularies, including the Random House® Unabridged Dictionary. As each word is selected its acoustic and language model statistics are immediately updated, transparent to the user. Once adapted to a particular speaker, the system can predict how that user will say new words that the system has never heard before.

The user doesn't have to be concerned with grammars, word lists, application dictionaries, or specifying pronunciations, even for new words. This knowledge is already built-in or acquired automatically during interactive system use. With its adaptive training, its performance and throughput continues to improve the more the system is used. Explicit training for individual words is also available on-line, and recommended for use with console command words. By using these console commands, the DragonDictate™ system can be operated entirely hands-free. Hands-free operation has proved especially valuable for both disabled users as well as non-disabled users desiring complete freedom from the keyboard.

USERS AND THEIR APPLICATIONS

The primary application for the DragonDictate™ system is interactive transcription of discretely dictated natural language free text. DragonDictate™ has been tested and may be used "as-is" with most popular ("well-behaved") MS-DOS word processing programs, as well as other standard "off-the-shelf" or custom database, spreadsheet, or report generation software. DragonDictate™ produced text may also be used for many mini and mainframe software packages via standard terminal emulation. In most cases, no special integration or application engineering is required to run DragonDictate™ with standard horizontal or vertical application software packages!

The DragonDictate™ system is especially oriented to professionals, business executives, and other people with limited typing skills. Hands-busy occupations, such as medicine, often preclude immediate key entry, regardless of an individual's typing abilities, or the pressing need to communicate vital information. Many people need to produce time-critical documents and reports, and do not have dedicated secretarial support. Present labor trends promise to exacerbate these problems yet further for the foreseeable future. Skilled typists, as typified by most able-bodied secretaries, computer programmers, and journalists, are not expected to derive significant benefits from "voice typewriters" until similar speech capabilities with continuous speech become commercially viable.

Although slower than with a skilled typist, most experienced DragonDictate™ users generate finished computer-readable printed text, simply by speaking, in less time than they could hand write the same words. Many people, (estimates run as high as 80 to 90% for professionals) presently hand write out their "discrete" text generation methodology! for someone else to type. Especially in environments where secretarial services are scarce, such individuals could now achieve real control over producing finished text to meet their own timetables. Medical, legal, and business communications are prime candidates. The economic value of timely turnaround for these and other documents such as proposals, and even sales quotes, is well known. This author believes that the broad appreciation of the business community for fast turnaround of important communications and self-sufficiency on the part of individuals creating these communications, has directly propelled the widespread success of photocopiers, word-processing, overnight courier services, and more recently, facsimile transmissions. Large vocabulary, natural language speech recognition has similar potential, although it will take more time and resources, and continuous speech input capabilities, to fully realize this potential.

Producing discrete speech is easy for most people within a few minutes of practice. The immediate display of the recognized speech in this paradigm, strongly facilitates a productive system interaction and positive user response. A brand new user typically will realize keystroke savings of 50%+ with an initial throughput of 15 wpm. After generating only 2 to 4 pages of text, these same users achieve keystroke savings often exceeding 70% with a throughput of 20 to 30 wpm. Well experienced users typically exceed 80 to 90%+ keystroke savings with a throughput ranging from 25 up to 60 wpm. Throughput is computed counting the number of words entered, recognized, and corrected where necessary, per unit of elapsed time. Throughput measures are computed on the same kind of task performed by skilled secretaries; namely, copying. For most professionals, the time consumed in creative text composition strongly dominates text creation times. That factor will not change even with the availability of continuous large vocabulary, natural language recognition. The productivity advantages of the DragonDictate™ system today are the faster generation of printed text for people writing time-sensitive documents long hand, especially those desiring more self-sufficiency and certainty in creating their finished text, when depending on uncertain secretarial resources.

DragonDictate™-30K is already being used operationally by people generating text for diverse applications, in standard working environments. Users include both naive and computer-wise individuals, and range in age from children to senior citizens. Applications include medical/legal reports, business documentation, database searches, personal correspondence, research reports, electronic messaging, computer programming, etc. Feedback from disabled users has been especially gratifying. DragonDictate™ is enabling these individuals to work far more effectively and become self-supporting. In the future we look forward to continuous speech capabilities and expect eventually that large vocabulary natural language, speech recognition will become the text generation vehicle of choice for most people, provided that it is well integrated into user applications, easy to use, and is economically justifiable. Initial responses from DragonDictate™ users and the market more generally, indicate that even at present, well integrated discrete natural language recognition, is probably providing a better solution for a significant population, than other alternatives presently available to them.

REFERENCE


ACKNOWLEDGEMENTS

It will presumably come as no surprise to the reader, that the text (word processing commands, formatting, etc.) for this paper has been entirely generated with spoken input to the DragonDictate™-30K transcription system.

The technological achievements culminating in DragonDictate™ represent many years of creative cooperative research and development on the part of an outstanding team of scientists / engineers at Dragon Systems, Inc. Significant contributions to DragonDictate™ have been made by James Baker, Janet Baker, Paul Bamberg, Laurence Gillick, David Pinto, Edward Porter, Jed Roberts, Robert Roth, Mark Sidell, and Dean Sturtevant.