Automated Dialing of Cellular Telephones Using Speech Recognition

Application Report

Frank Henry Dearden III
Voice Control Systems, Incorporated

SPRA144
October 1994

PRINTED WITH SOY INK™

TEXAS INSTRUMENTS

Printed on Recycled Paper

IMPORTANT NOTICE

Texas Instruments (TI) reserves the right to make changes to its products or to discontinue any semiconductor product or service without notice, and advises its customers to obtain the latest version of relevant information to verify, before placing orders, that the information being relied on is current.

TI warrants performance of its semiconductor products and related software to the specifications applicable at the time of sale in accordance with TI's standard warranty. Testing and other quality control techniques are utilized to the extent TI deems necessary to support this warranty. Specific testing of all parameters of each device is not necessarily performed, except those mandated by government requirements.

Certain applications using semiconductor products may involve potential risks of death, personal injury, or severe property or environmental damage ("Critical Applications").

TI SEMICONDUCTOR PRODUCTS ARE NOT DESIGNED, INTENDED, AUTHORIZED, OR WARRANTED TO BE SUITABLE FOR USE IN LIFE-SUPPORT APPLICATIONS, DEVICES OR SYSTEMS OR OTHER CRITICAL APPLICATIONS.

Inclusion of TI products in such applications is understood to be fully at the risk of the customer. Use of TI products in such applications requires the written approval of an appropriate TI officer. Questions concerning potential risk applications should be directed to TI through a local SC sales office.

In order to minimize risks associated with the customer's applications, adequate design and operating safeguards should be provided by the customer to minimize inherent or procedural hazards.

TI assumes no liability for applications assistance, customer product design, software performance, or infringement of patents or services described herein. Nor does TI warrant or represent that any license, either express or implied, is granted under any patent right, copyright, mask work right, or other intellectual property right of TI covering or relating to any combination, machine, or process in which such semiconductor products or services might be or are used.

Introduction

The cellular telephone industry has experienced tremendous growth since its beginning more than ten years ago. What was once considered to be a toy for high-profile executives has now become an integral communications tool for over 14 million subscribers in the U.S. alone. Growth rates are expected to accelerate during the next few years.

Automated speech recognition (ASR) technology has been a bedfellow of cellular telephone technology for many years. Most of the large cellular subscriber unit manufacturers have developed their own ASR systems to facilitate hands-free dialing. The benefits of combining these two technologies are obvious: the less time and focus a driver gives to placing a call, the more attentive he is to operating the vehicle. Hands-free kits that include a far-talk microphone and speaker are now required by law in some European countries for conversing once a call is connected. Various states are currently considering similar requirements. Similarly, requirements for hands-free dialing capability via speech recognition are not too far off.

This paper explains how ASR-enabled dialing capability can be implemented with DSP technology from Texas Instruments. Speech recognition technology has never been as accurate, user-friendly, and inexpensive as it is today, or as easy to integrate into state-of-the-art cellular subscriber systems.

The Technology

Most of the past and existing ASR units on the market are limited to what is known as *speaker-dependent* (SD) technology. This technology has exhibited some rather fundamental performance limitations. SD systems work by comparing a whole word input with a user-supplied *template*. Templates are developed by each user during a rather cumbersome training exercise, which usually takes place in a quiet, stationary environment. Since the systems are used in a moving car environment, the increase in background noise, coupled with a user's inflection change (people usually shout slightly, and unconsciously, when a car is in motion) confuse most SD systems. Accuracy rates are typically less than 90%.

Speaker-dependent ASR systems are steadily being replaced with *speaker-independent* (SI) systems. SI-capable systems approach the recognition problem in a fundamentally different manner than SD-only systems. Once an input command is captured and digitized, an SI system will parse it into phonetic-like pieces, or *features*. These speech features are then compared with supplied target data, not with templates supplied by the user.

The training procedure for a speaker-independent recognizer is both processing and data intensive. Speech variations due to sex, age, accent, and speaking habits must be considered, along with the great variety of noise sources, internal and external to the car, that have a tremendous effect on the signal-to-noise ratio. This implies that an application-specific speech data base is required for the vocabulary *training* process. Consequently, each SI vocabulary is essentially hand-crafted for the particular word list and the environment of use. The diversity of the training data helps account for the robustness of the resultant recognizer in the presence of real users and all types of automobile noise.

Usually, speech-independent reference data is derived from a large data base of speech *tokens* collected inside several cars, from hundreds of speakers, over a variety of road conditions, and with high-quality digital recording equipment. The computer-controlled recording equipment has a display screen that automatically prompts the *donor* to speak through a given vocabulary. The incoming speech sample is transduced by a noise-canceling microphone placed on the windshield and is recorded on a remotely controlled digital audio tape (DAT).

The result is a scheme that is extremely robust. Matching pieces of sound to feature templates derived from rigorously collected data reduces the amount of computational power required and is more forgiving of inflection change than an SD scheme. For example, a cold will make John sound less like John specifically, but his speech will continue to exhibit feature characteristics consistent with the statistical samples derived from the database. Additionally, technologists at Voice Control Systems Incorporated (VCS) have done empirical analysis on SI feature recognition and have even identified some features that occur often but are irrelevant to recognition. The complexity of the task can be reduced and the odds of a successful recognition increased if some of these redundant features are disregarded.

The Human Interface

All recognition systems consist of two basic components: the core recognition engine and the human interface. The adage "you're only as good as your presentation" is very apropos when designing ASR systems. Technologists tend to devote most of their time to enhancing a system's raw recognition power, bandwidth, and memory allocation, etc. This is all well and good. Marketers however, should make sure that the interface gets an equal amount of attention.

Besides high accuracy, the major benefit of a speaker-independent capable system is its intuitive, user-friendly presentation. Acceptance by the user is critical, especially during the first use. The system should prompt the user with high-quality, stored human speech and should respond quickly to each input. The result should be a semiconversational experience, such as the following (the user input is in bold **CAPITALS** and the response is in lowercase):

"VOICE CONTROL"
"ready"
"CALL"
"calling?"
"OFFICE"
"calling office, correct?"
"YES"
"dialing..."

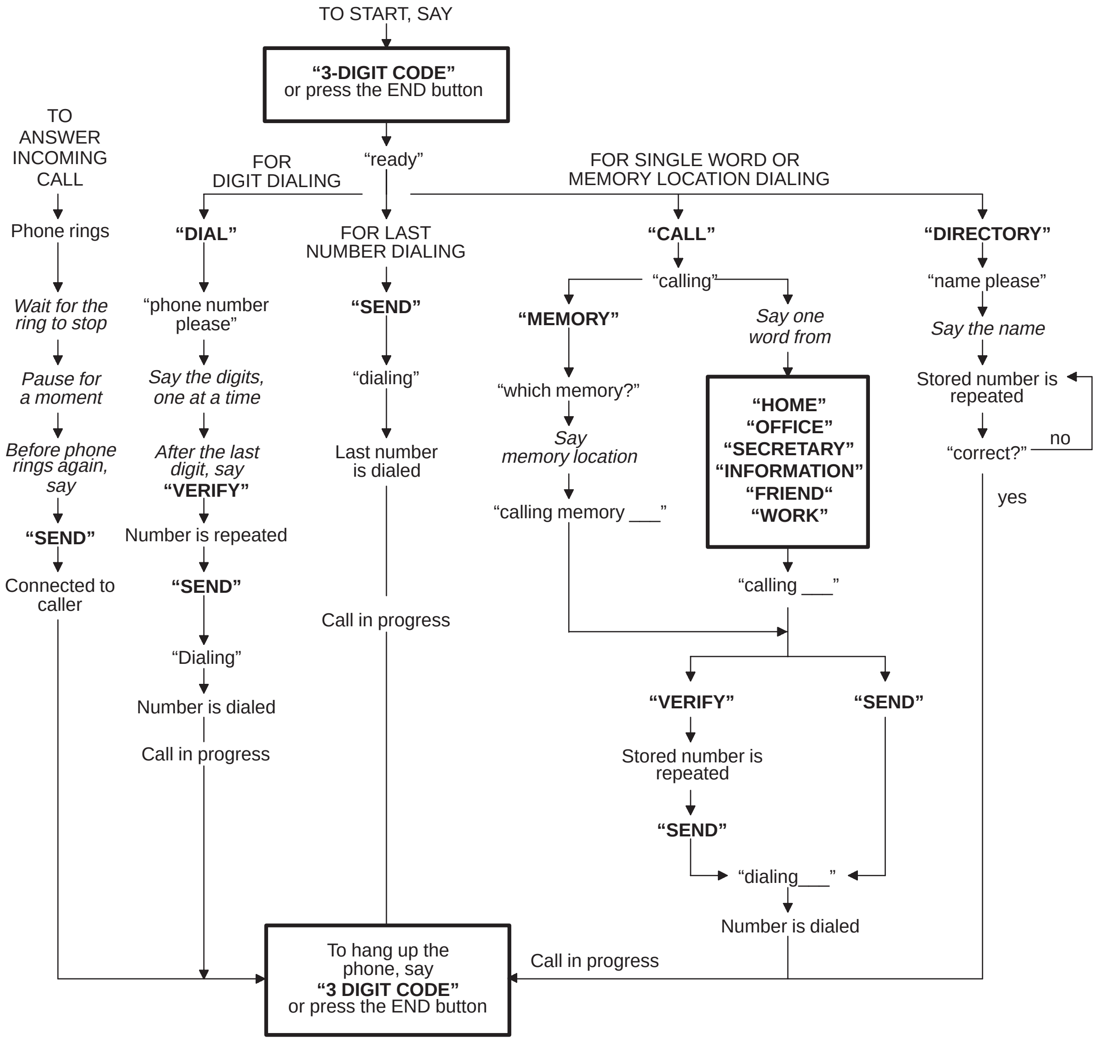
In this example, the user accesses a memory location by using one of many possible predetermined name tags (for example, office, home, school, information, doctor, etc.).

A user should also be able to place a random phone call by using a speaker-independent digit dialing sequence, like this:

"DIAL"
"phone number, please"
"THREE"
[beep, display 3]
"SEVEN"
[beep, display 7]
etc.
"VERIFY"
"three, seven, (etc.)"
"SEND"
"dialing..."

Figure 1 shows a flowchart, or decision tree, of a well-tested human interface.

Figure 1. Flow Diagram of Human Interface



NOTE: User input is in bold **CAPITALS**, the response is in lowercase, and directions are in *italics*.

Note that a system can be both speaker-independent capable and speaker-dependent capable. SD technology allows a user to assign personal name tags to memory locations in addition to the SI locations mentioned above. Depending on the memory available, a user can program phone numbers into memory locations labeled "John Smith", "Fred's Office", "Pizza", etc. For the greatest recognition accuracy, it is best to limit the number of customizable name tags to about ten. VCS uses its feature-matching algorithm for SD comparisons as well as for SI comparisons, resulting in high accuracy rates.

The Implementation

VCS has focused solely on developing ASR technology for the past 14 years. Most of VCS's more than 90,000 fielded systems are multichannel telephone network-based installations, which allow random callers to utilize voice mail or other interactive response functions without the need for touch-tone input. The recognition algorithms in these applications are handled by dedicated TI DSP hardware.

VCS began to migrate its ASR expertise into single-channel applications about four years ago. The goal was to maintain high functionality while minimizing hardware cost and space requirements. These first products used custom interface circuitry, standard X86 microprocessors, a CODEC, and memory for the core recognition hardware. The custom chip has been redesigned to reduce cost for the recognition core to about $30 in quantities of 10,000. Manufacturing tooling, testing, packaging, and labor expenses can easily lead to a total cost per unit of twice this amount. Although the circuit can be made quite small, adequate space must be allowed in a transceiver unit, a 3-watt booster, an external enclosure, or even within a handset cradle. Building this chipset directly into a portable cellular telephone remains impractical at this time.

ASR can also be used with digital cellular phones because VCS code can take advantage of the hardware already resident within the handset. This hardware includes a CODEC, memory, and digital signal processing capability. Consequently, adding ASR code may require some additional memory capacity but does not require the design and manufacture of an entire circuit board. The total cost is greately reduced—from about $60 to about $15—which includes additional memory and software licensing.

Since VCS's ASR code uses only a small amount of DSP bandwidth, the cellular telephone can execute recognition operations in parallel with being enabled for incoming calls. For example, our discrete speaker-independent and speaker-dependent capability utilizes about 25% of the bandwidth for one channel of recognition on a TMS320C25 operating at 40 MHz. Additionally, the ASR code does not compete with the digitization and companding exercises undertaken during a conversation, because the recognition task for dialing precedes the actual placement of a call.

In this scheme, the cellular telephone task master communicates with VCS ASR object code via applications programming interface (API) commands. This involves a reasonable level of integration, but the end result is the lowest incremental cost option for adding ASR to a cellular telephone. An API for the Texas Instruments TMS320C2x DSPs can be acquired directly from VCS.

Accuracy

VCS has designed a tape test exercise to systematically determine the recognition accuracies of a newly designed voice recognition unit (VRU); the procedures for quantifying the performance of speaker-independent and speaker-dependent commands are different. A properly designed VRU will utilize these two technologies to maximize the acceptability of the system by the operator.

Tape testing is conducted under laboratory conditions and with a direct audio path between the tape and the VRU. The total number of SI commands a system is capable of recognizing is simply a function of available memory. However, at any given time, only a specific subset of the total SI vocabulary should be active. In general, each subvocabulary should be limited to about 12 elements, even though larger subvocabularies are possible. Smaller subvocabularies maximize the performance of the technology and minimize operator choice and confusion. Each speaker-independent subvocabulary (that is, each path in the *tree*) should be tested.

The test data includes 50 speakers, of which half are male and half are female. The data is obtained from a data collection of every recognizable word in the vocabulary, as described above. These data are reserved for testing purposes only and are not to be used to train the VRU.

Each response is recorded as the source tape is played. Twice, the tape plays each person speaking the entire speaker-independent vocabulary, divided into the designated subvocabularies. The expected error rates for VCS speaker-independent technology are:

Average rejection error rate	< 3.0%
Average substitution error rate	< 1.5%

A rejection error occurs when the system rejects a valid word input on the basis of insufficient *class* distinction. A substitution error occurs when the system substitutes another word from the active vocabulary in response to a valid word input. On occasion (less than 1% of the time), the system may not respond to a spoken input, because the word was not spoken loud enough. These cases should be ignored when the rejection and substitution error rates are computed.

Softening the impact of an error is the job of the user-friendly interface. For example, if the VRU responds with a polite *"pardon?"* following a rejection error, most people will patiently repeat the input (at least once) and enunciate a bit more clearly. The system typically accepts the next attempt, and the user proceeds, sometimes unaware that an error has occurred. For this reason, an SI rejection error rate under 4% is perfectly acceptable for most users.

VCS systems have the capability to handle at least one SD vocabulary, although with enough memory, more are possible. However, only one vocabulary should be active at any given time. During testing, this speaker-dependent memory should initially be cleared. A representative group of ten people, five male and five female, should participate, with a minimum of three passes. Words not easily confused should be used for this test.

home	office	Steve	Bob	Mary Jones
Sears	Jill Miller	weather	voice mail	John Smith

Each member of the group then rotates through the above list ten times, trying to recall the correct command. On average, the expected substitution error rates for VCS speaker-dependent vocabularies are less than 5%. SD vocabularies are not prone to rejection errors.

It is extremely difficult to combine technologies, (that is, to have a speaker-independent vocabulary simultaneously active with a speaker-dependent vocabulary). Situations like this should be avoided, if for no other reason than to minimize the confusion of the operator.

Code Availability

The associated software is available for licensing from Voice Control Systems, 14140 Midway Road, Dallas, Texas 75244. Relevant data sheets are also included in the *TMS320 Software Cooperative Data Sheet Folder,* Texas Instruments literature number SPRT111.

Summary

With a PC, multimedia hardware, and a relevant technical paper in the public domain, an engineer can design a reasonable speaker-dependent ASR system. The accuracy is usually in the mid-80% range, as long as the environment is quiet. Improving this capability to handle speaker-independent input, achieve a 97%+ accuracy in noisy environments, and cost as little as $15 per unit is quite another challenge.

VCS has worked for more than a decade in tedious research and testing to incrementally improve its technology to these levels. It is predicted that the features and benefits offered by ASR will greatly influence subscriber unit purchases.