APPLICATIONS OF SPEECH RECOGNITION IN THE AREA OF TELECOMMUNICATIONS

Lawrence R. Rabiner
AT&T Labs
Florham Park, New Jersey 07932

Abstract – Advances in speech recognition technology, over the past 4 decades, have enabled a wide range of telecommunications and desktop services to become 'voice-enabled'. Early applications were driven by the need to automate and thereby reduce the cost of attendant services, or by the need to create revenue generating new services which were previously unavailable because of cost, or the inability to adequately provide such a service with the available work force. As we move towards the future we see a new generation of voice-enabled service offerings emerging including intelligent agents, customer care wizards, call center automated attendants, voice access to universal directories and registries, unconstrained dictation capability, and finally unconstrained language translation capability. In this paper we review the current capabilities of speech recognition systems, show how they have been exploited in today's services and applications, and show how they will evolve over time to the next generation of voice-enabled services.

1. INTRODUCTION

Speech recognition technology has evolved for more than 40 years, spurred on by advances in signal processing, algorithms, architectures, and hardware. During that time it has gone from a laboratory curiosity, to an art, and eventually to a full fledged technology that is practiced and understood by a wide range of engineers, scientists, linguists, psychologists, and systems designers. Over those 4 decades the technology of speech recognition has evolved, leading to a steady stream of increasingly more difficult tasks which have been tackled and solved. The hierarchy of speech recognition problems which have been attacked, and the resulting application tasks which became viable as a result, includes the following [1]:

- *isolated word recognition*-both speaker trained and speaker independent. This technology opened up a class of applications called 'command-and-control' applications in which the system was capable of recognizing a single word command (from a small vocabulary of single word commands), and appropriately responding to the recognized command. One key problem with this technology was the sensitivity to background noises (which were often recognized as spurious spoken words) and extraneous speech which was inadvertently spoken along with the command word. Various types of 'keyword spotting' algorithms evolved to solve these types of problems.
- *connected word recognition*-both speaker trained and speaker independent. This technology was built on top of word recognition technology, choosing to exploit the word models that were successful in isolated word recognition,

0-7803-3698-4/97/$10.00 © 1997 IEEE.

and extend the modeling to recognize a concatenated sequence (a string) of such word models as a word string. This technology opened up a class of applications based on recognizing digit strings and alphanumeric strings, and led to a variety of systems for voice dialing, credit card authorization, directory assistance lookups, and catalog ordering.

- *continuous or fluent speech recognition*-both speaker trained and speaker independent. This technology led to the first large vocabulary recognition systems which were used to access databases (the DARPA Resource Management Task), to do constrained dialogue access to information (the DARPA ATIS Task), to handle very large vocabulary read speech for dictation (the DARPA NAB Task), and eventually were used for desktop dictation systems for PC environments [2].
- *speech understanding systems* (so-called unconstrained dialogue systems) which are capable of determining the underlying message embedded within the speech, rather than just recognizing the spoken words [3]. Such systems, which are only beginning to appear recently, enable services like customer care (the AT&T How May I Help You System), and intelligent agent systems which provide access to information sources by voice dialogues (the AT&T Maxwell Task).
- *spontaneous conversation systems* which are able to both recognize the spoken material accurately and understand the meaning of the spoken material. Such systems, which are currently beyond the limits of the existing technology, will enable new services such as 'Conversation Summarization', 'Business Meeting Notes', 'Topic Spotting' in fluent speech (e.g., from radio or TV broadcasts), and ultimately even language translation services between any pair of existing languages.

1.1 Generic Speech Recognition System [4]

Figure 1 shows a block diagram of a typical integrated continuous speech recognition system. Interestingly enough, this generic block diagram can be made to work on virtually any speech recognition task that has been devised in the past 40 years, i.e., isolated word recognition, connected word recognition, continuous speech recognition, etc.

The feature analysis module provides the acoustic feature vectors used to characterize the spectral properties of the time varying speech signal. The word-level acoustic match module evaluates the similarity between the input feature vector sequence (corresponding to a portion of the input speech) and a set of acoustic word models for all words in the recognition task vocabulary to determine which words were most likely spoken. The sentence-level match module uses a language model (i.e., a model of syntax and semantics) to determine the most likely sequence of words. Syntactic and semantic rules can be specified, either manually, based on task constraints, or with statistical models such as word and class N-gram probabilities. Search and recognition decisions are made by considering all likely word sequences and choosing the one with the best matching score as the recognized sentence.

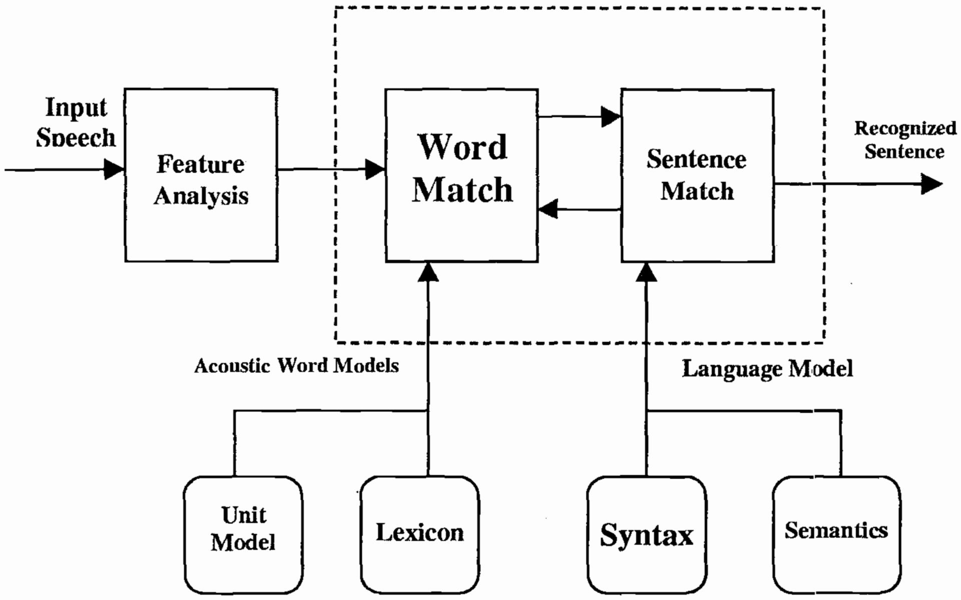

Figure 1 Block diagram of a typical integrated continuous speech recognizer.

Almost every aspect of the continuous speech recognizer of Figure 1 has been studied and optimized over the years. As a result, we have obtained a great deal of knowledge about how to design the feature analysis module, how to choose appropriate recognition units, how to populate the word lexicon, how to build acoustic word models, how to model language syntax and semantics, how to decode word matches against word models, how to efficiently determine a sentence match, and finally how to eventually choose the best recognized sentence. Among the things we have learned are the following:

- the best spectral features to use are LPC-based cepstral coefficients (either on a linear or a mel frequency scale) and their first and second order derivatives, along with log energies and their derivatives.
- the continuous density hidden Markov model (HMM) with state mixture densities is the best model for the statistical properties of the spectral features over time.
- the most robust set of speech units is a set of context dependent triphone units for modeling both intraword and interword linguistic phenonema.
- although maximum likelihood training of unit models is effective for many speech vocabularies, the use of discriminative training methods (e.g., MMI training or Global Probabilistic Descent (GPD) methods) is more effective for most tasks.

- the most effective technique for making the unit models be robust to varying speakers, microphones, backgrounds, and transmission environments is through the use of signal conditioning methods such as Cepstral Mean Subtraction (CMS) or some type of Signal Bias Removal (SBR).
- the use of adaptive learning increases performance for new talkers, new backgrounds, and new transmission systems.
- the use of utterance verification provides improved rejection of improper speech or background sounds.
- HMM's can be made very efficient in terms of computation speed, memory size, and performance through the use of subspace and parameter tieing methods.
- efficient word and sentence matches can be obtained through the use of efficient beam searches, tree-trellis coding methods, and through proper determinization of the Finite State Network (FSN) that is being searched and decoded. Such procedures also lead to efficient methods for obtaining the *N*-best sentence matches to the spoken input.
- the ideas of concept spotting can be used to implement semantic constraints of a task in an automatically trainable manner.

1.2 Building Good Speech-Based Applications [5]

In addition to having good speech recognition technology, effective speech-based applications heavily depend on several factors, including:

- good user interfaces which make the application easy-to-use and robust to the kinds of confusion that arise in human-machine communications by voice.
- good models of dialogue that keep the conversation moving forward, even in periods of great uncertainty on the parts of either the user or the machine.
- matching the task to the technology.

We now expand somewhat on each of these factors.

User Interface Design-In order to make a speech interface as simple and as effective as Graphical User Interfaces (GUI), 3 key design principles should be followed as closely as possible, namely:

- provide a *continuous representation* of the objects and actions of interest.
- provide a mechanism for *rapid, incremental, and reversible* operations whose impact on the object of interest is immediately visible.
- use physical actions or labeled button presses instead of text commands, whenever possible.

For Speech Interfaces (SI), these GUI principles are preserved in the following user design principles:

- remind/teach users what can be said at any point in the interaction.
- maintain consistency across features using a vocabulary that is 'almost always available'.
- design for error.
- provide the ability to barge-in over prompts.
- use implicit confirmation of voice input.

- rely on 'earcons' to orient users as to where they are in an interaction with the machine.
- avoid information overload by aggregation or pre-selection of a subset of the material to be presented.

These user interface design principles are applied in different ways in the applications described later in this paper.

Dialogue Design Principles-For many interactions between a person and a machine, a dialogue is needed to establish a complete interaction with the machine. The 'ideal' dialogue allows either the user or the machine to initiate queries, or to choose to respond to queries initiated by the other side. (Such systems are called 'mixed initiative' systems.) A complete set of design principles for dialogue systems has not yet evolved (it is far too early yet). However, much as we have learned good speech interface design principles, many of the same or similar principles are evolving for dialogue management. The key principles that have evolved are the following:

- summarize actions to be taken, whenever possible.
- provide real-time, low delay, responses from the machine and allow the user to barge-in at any time.
- orient users to their 'location' in task space as often as possible.
- use flexible grammars to provide incrementality of the dialogue.
- whenever possible, customize and personalize the dialogue (novice/expert modes).

In addition to these design principles, an objective performance measure is needed that combines task-based success measures (e.g., information elements that are correctly obtained) and a variety of dialogue-based cost measures (e.g., number of error correction turns, time to task completion, success rate, etc.) Such a performance measure for dialogues does not yet exist but is under investigation.

Match Task to the Technology-It is essential that any application of speech recognition be realistic about the capabilities of the technology, and build in failure correction modes. Hence building a credit card recognition application before digit error rates fell below 0.5% per digit is a formula for failure, since for a 16-digit credit card, the string error rate will be at the 10% level or higher, thereby frustrating customers who speak clearly and distinctly, and making the system totally unusable for customers who slur their speech or otherwise make it difficult to understand their spoken inputs. Utilizing this principle, the following successful applications have been built:

- telecommunications: Command-and-Control, agents, call center automation, customer care, voice calling.
- office/desktop: voice navigation of desktop, voice browser for Internet, voice dialer, dictation.
- manufacturing/business: package sorting, data entry, form filling.
- medical/legal: creation of stylized reports.

- games/aids-to-the-handicapped: voice control of selective features of the game, the wheel chair, the environment (climate control).

1.3 Current Capabilities of Speech Recognizers

Table 1 provides a summary of the performance of modern speech recognition and natural language understanding systems. Shown in the table are the Task or Corpus, the Type of speech input, the Vocabulary Size and the resulting Word Error Rate. It can be seen that the technology is more than suitable for connected digit recognition tasks, for simple data retrieval tasks (like the Airline Travel Information System), and, with a well-designed user interface, can even be used for dictation like the Wall Street Journal Task. However, the word error rates rapidly become prohibitive for tasks like recognizing speech from a radio broadcast (with all of the cross-announcer banter, commercials, etc), from listening in on conversational telephone calls off a switchboard, or even in the case of familiarity of families calling each other over a switched telephone line.

CORPUS	*TYPE*	*VOCABULARY SIZE*	*WORD ERROR RATE*
Connected Digit Strings	Spontaneous	10	0.3%
Airline Travel Information	Spontaneous	2500	2.0%
Wall Street Journal	Read Text	64,000	8.0%
Radio (Marketplace)	Mixed	64,000	27%
Switchboard	Conversational Telephone	10,000	38%
Call Home	Conversational Telephone	10,000	50%

Table 1 Word Error Rates for Speech Recognition and Natural Language Understanding Tasks (Courtesy: John Makhoul, BBN)

1.4 Instantiations of Speech Recognition Technology

Speech recognition technology used to be available only on special purpose boards with special purpose DSP chips or ASIC's. Today high quality speech recognition technology packages are available in the form of inexpensive software only desktop packages (IBM ViaVoice, Dragon Naturally Speaking, Kurzweil, etc.), technology engines that run on either the desktop or a workstation and are often embedded in third party vendor applications, such as the BBN Hark System, the SRI Nuance System, the AT&T Watson System, and the Altech System, and finally they are also available as proprietary engines running on commercially available speech processing boards such as the Lucent Speech Processing System (LSPS), the TI board, the Nortel board, etc.

2. The Telecommunications Need for Speech Recognition [6]

The telecommunications network is evolving as the traditional POTS (Plain Old Telephony Services) network comes together with the dynamically evolving Packet network, in a structure which we believe will look something like the one shown in Figure 2.

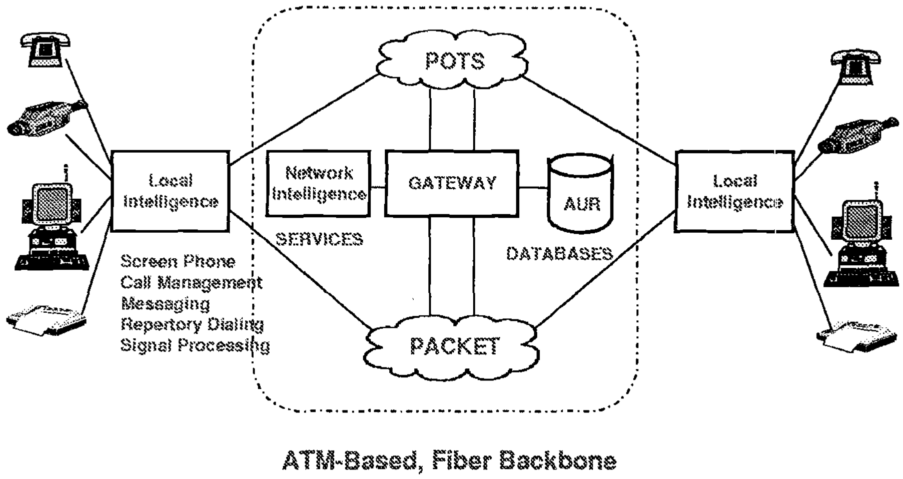

Figure 2 The telecommunications network of tomorrow.

Intelligence in this evolving network is distributed at the desktop (the local intelligence), at the terminal device (the telephone, screen phone, PC, etc), and in the network. In order to provide universal services, there needs to be interfaces which operate effectively for all terminal devices. Since the most ubiquitous terminal device is still the ordinary telephone handset, the evolving network must rely on the availability of speech interfaces to all services. Hence the growing need for speech recognition for Command-and-Control applications, and natural language understanding for maintaining a dialogue with the machine.

3. Telecommunication Applications of Speech Recognition [7]

Speech recognition was introduced into the telecommunications network in the early 1990's for two reasons, namely to reduce costs via automation of attendant functions, and to provide new revenue generating services that were previously impractical because of the associated costs of using attendants.

Examples of telecommunications services which were created to achieve cost reduction include the following:

- *Automation of Operator Services.* Systems like the Voice Recognition Call Processing (VRCP) system introduced by AT&T or the Automated Alternate Billing System (AABS) introduced by Nortel enabled operator functions to be handled by speech recognition systems. The VRCP system handled so-called 'operator assisted' calls such as Collect, Third Party Billing, Person-to-Person, Operator Assisted Calling, and Calling Card calls. The AABS system automated the acceptance (or rejection) of billing charges for reverse calls by recognizing simple variants of the two word vocabulary Yes and No.
- *Automation of Directory Assistance.* Systems were created for assisting operators with the task of determining telephone numbers in response to customer queries by voice. Both NYNEX and Nortel introduced a system that did front end city name recognition so as to reduce the operator search space for the desired listing, and several experimental systems were created to complete the directory assistance task by attempting to recognize individual names in a directory of as many as 1 million names. Such systems are not yet practical (because of the confusability among names) but for small directories, such systems have been widely used (e.g., in corporate environments).
- *Voice Dialing.* Systems have been created for voice dialing by name (so-called alias dialing such as Call Home, Call Office) from AT&T, NYNEX, and Bell Atlantic, and by number (AT&T SDN/NRA) to enable customers to complete calls without having to push buttons associated with the telephone number being called.

Examples of telecommunications services which were created to generate new revenue include the following:

- *Voice Banking Services.* A system for providing access to customer accounts, account balances, customer transactions, etc. was first created in Japan by NTT (the ANSER System) more than 10 years ago in order to provide a service that was previously unavailable. Equivalent services have been introduced in banks worldwide over the last several years.
- *Voice Prompter.* A system for providing voice replacement of touch-tone input for so-called Interactive Voice Response (IVR) systems was introduced by AT&T in the early 1990's (initially in Spain because of the lack of touch-tone phones in that country). This system initially enabled the customer to speak the touch-tone position (i.e., speak or press the digit one); over time systems have evolved so that customers can speak the service associated with the touch-tone position (e.g., say reservations or push the 1-key, say schedule or push the 2-key, etc.).
- *Directory Assistance Call Completion.* This system was introduced by both AT&T and NYNEX to handle completion of calls made via requests for Directory Assistance. Since Directory Assistance numbers are provided by an independent system, using Text-to-Speech synthesis to speak out the listing,

speech recognition can be used to reliably recognize the listing and dial the associated number. This highly unusual use of a speech recognizer to interface with a speech synthesizer is one of the unusual outgrowths of the fractionation of the telephone network into local and long distance carriers in the United States.

- *Reverse Directory Assistance.* This system was created by NYNEX, Bellcore, and Ameritech to provide name and address information associated with a spoken telephone number.
- *Information Services.* These type of systems enable customers to access information lines to retrieve information about scores of sporting events, traffic reports, weather reports, theatre bookings, restaurant reservations, etc.

As we move to the future the intelligent network of Figure 2, along with advances in speech recognition technology, will support a new range of services of the following types:

- *Agent Technology.* Systems like Wildfire and Maxwell (AT&T) enable customers to interact with intelligent agents via voice dialogues in order to manage calls (both in-coming and out-going calls), manage messages (both voice and email), get information from the Web (e.g., movie reviews, calling directories), customize services (e.g., first thing each morning the agent provides the traffic and weather reports), personalize services (via the agent personality, speed, helpfulness), and adapt to user preferences (e.g., learn how the user likes to do things and react appropriately).
- *Customer Care.* The goal of customer care systems is to replace Interactive Voice Response systems with a dialogue type of interaction to make it easier for the user to get the desired help without having to navigate complicated menus or understand the terminology of the place being called for help. The How May I Help You (HMIHY) customer care system of AT&T is an excellent example of this type of system.
- *Computer-Telephony Integration.* Since the telecommunication network of the future will integrate the telephony (POTS) and computer (Packet) networks, a range of new applications will arise which exploit this integration more fully. One prime example is registry services where the network locates the user and determines the most appropriate way to communicate with them. Another example is providing a user cache of the most frequently accessed people in order to provide a rapid access mechanism for these frequently called numbers.
- *Voice Dictation.* Although the desktop already supports voice dictation of documents, a prime telecommunications application of speech recognition would be for generating voice responses to email queries so that the resulting message becomes an email message back to the sender (rather than a voice mail response to an email message).

4. Summary

The world of telecommunications is rapidly changing and evolving. The world of speech recognition is rapidly changing and evolving. Early applications of the technology have achieved varying degrees of success. The promise for the future is significantly higher performance for almost every speech recognition technology area, with more robustness to speakers, background noises etc. This will ultimately lead to reliable, robust voice interfaces to every telecommunications service that is offered, thereby making them universally available.

References

[1] L. R. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition*, Englewood Cliffs, NJ, 1993.

[2] J. Makhoul and R. Schwartz, "State of the Art in Continuous Speech Recognition", in *Voice Communications Between Humans and Machines*, D. Roe and J. Wilpon, Eds., pp. 165-198, 1994.

[3] R. Pieraccini and E. Levin, "Stochastic Representation of Semantic Structure for Speech Understanding", *Speech Communications*, Vol. 11, pp. 283-288, 1992.

[4] L. R. Rabiner, B. H. Juang, and C. H. Lee, "An Overview of Automatic Speech Recognition", in *Automatic Speech and Speaker Recognition*, C. H. Lee, F. K. Soong, and K. K. Paliwal, Eds., pp. 1-30, 1996.

[5] C. A. Kamm, M. Walker, and L. R. Rabiner, "The Role of Speech Processing in Human-Computer Intelligent Communication", *Proc. HCI Workshop*, Washington, DC, pp. 169-190, Feb. 1997.

[6] R. V. Cox, B. G. Haskell, Y. LeCun, B. Shahraray, and L. R. Rabiner, "On the Applications of Multimedia Processing to Communications", submitted to *Proc. IEEE.*

[7] L. R. Rabiner, "Applications of Voice Processing to Telecommunications", *Proc. IEEE*, Vol. 82, No. 4, pp. 199-228, Feb. 1994.