

Mazin Gilbert, Jay G. Wilpon, Benjamin Stern,
and Giuseppe Di Fabbrizio

[A human-machine communication
system for next-generation
contact centers]



© ARTVILLE & COMSTOCK

Intelligent Virtual Agents for Contact Center Automation

The evolution of communication services over the past century has spawned a broad new industry known as electronic contact (eContact), which provides electronic communication mechanisms between people and businesses or organizations. One byproduct of this evolution, led by advancements in speech and language technologies, is enabling people to engage in seamless, natural conversation with a new breed of intelligent services. These services, which act as virtual agents, are capable of recognizing and understanding natural conversational speech. They learn from data, can scale with minimal human intervention, and are able to extract intelligence information to help improve their operation and the business they serve. This article highlights some of AT&T's technical innovations within VoiceTone, the subset of the eContact space focused on creating intelligent virtual agents for contact center automation.

Over the past century, the eContact industry has been slowly emerging, providing electronic communication between people (e.g., customers, employees, suppliers) and organizations (e.g., businesses, enterprises, government agencies) to accomplish or facilitate a business transaction. It has been historically linked with advances in communication technology. It encompasses multiple types of media (speech, Web sites, e-mails, chat, and video) over multiple channels (telephone network, wireless, and broadband Internet). One byproduct of this technology advancement is the evolution of the user interface, as shown in Figure 1. The voice-to-voice interface, a natural user experience, eventually has given way to a human-to-machine

interface, first via touch-tone prompts and ultimately via today's multimedia and multimodal eContact.

This new wave of technological advancements promises to provide ubiquitous and personalized global access to communication services. It will ultimately constitute a vehicle to automate services without agent personnel. It will be particularly valuable for the contact center market, which is currently estimated to cost U.S. industry in excess of US\$100 billion annually. Of this total, 80% is spent on agent personnel, including direct expenses (salaries) and indirect expenses (benefits, administration, and buildings). Automating even a fraction of the calls currently handled by these agents will ultimately generate a tremendous cost savings.

This article highlights some of the technical innovations within AT&T's service enterprise, VoiceTone, the subset of the eContact space focused on voice-enabled contact center automation. VoiceTone comprises all the media, linkages, and controls between a person (with a request for information or service) and the business delivery system that can satisfy that request (including a human agent). It uses advances in speech and language technology to enable customers to access and manipulate information by conversing naturally with an intelligent virtual agent. The goal is not only to help contact centers reduce their operating costs and increase their service revenue but also to improve customer experience over existing interactive voice response (IVR) systems. AT&T VoiceTone applications are based on systematically mixing both open (or user-initiative) and directed (or system-initiative) dialog strategies to provide users the ability to speak naturally and complete their transaction requests. They represent a new generation of automated contact center services that greet the customer with the open-ended prompt "How may I help you?" In this way, callers can describe their request naturally using their own words, as opposed to the words dictated to them by the machine. The dialog technology enables a proper balance between the system-controlled and the user-controlled strategies to best meet the business requirements for complete fulfillment of transactions.

AT&T VOICETONE

The VoiceTone platform is based on a typical three-tier Web architecture with a clear separation among the client (with VoiceXML browser [1]), the application server hosting the application logic [2], and the enterprise information system or database backend. Besides supporting standardized interfaces to promote more plug-n-play components, this architecture is designed to enable users to engage in an open, natural-language dialog. The notion of a cumbersome, menu-based interaction offered by a traditional directed dialog approach is replaced by a natural, flexible user experience. Open-dialog-based services shift the burden of

usage from the user to the system and associated speech technologies. This simulates human-to-human-like interactions, replacing the cumbersome IVR-like experience that directed dialog services tend to impose. The following interaction is an example of a VoiceTone open dialog service:

VoiceTone: Hello, this is Barney Health Care. How may I help you?

User: My medicine ran out last week.

VoiceTone: Would you like to refill your prescription?

User: yep, uh, my refill I.D. number is PB14720.

VoiceTone: Okay, I have ordered a new prescription for you. Is there anything else I can help you with?

User: No thanks.

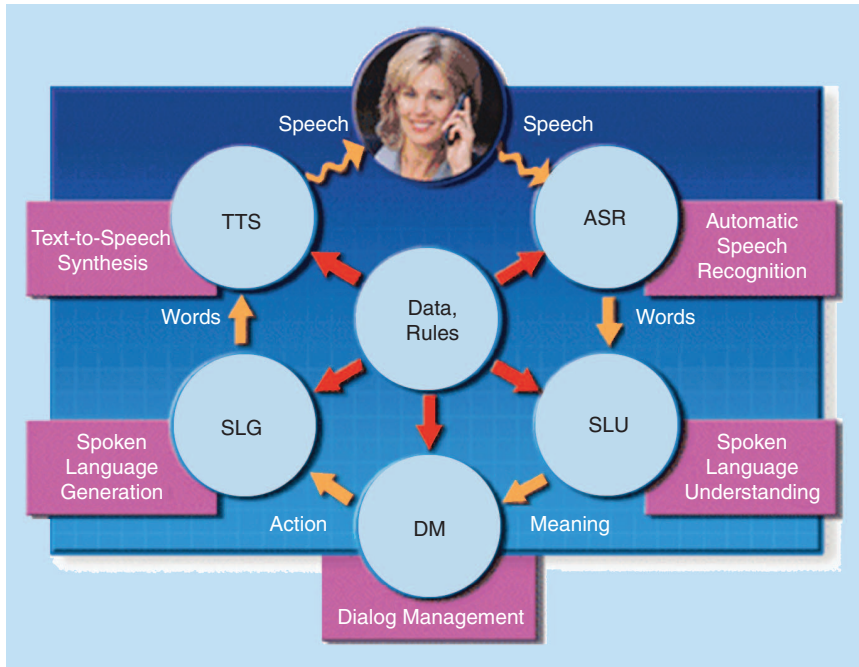
VoiceTone: Thank you for calling Barney Health Care. Good-bye.

The main technologies behind open dialog services (shown in Figure 2) include automatic speech recognition (ASR), spoken language understanding (SLU), dialog management (DM), spoken language generation (SLG), and text-to-speech (TTS) synthesis. Each of these technologies is effectively acting as a transducer. ASR converts a speech signal into a sequence (or a lattice) of words. SLU processes this lattice and extracts the meaning. The DM uses the meaning as well as dialog history to generate an action that is converted into words by the SLG. The TTS converts the output words and any available speech tags into synthesized speech. This cycle is equivalent to one turn in a human/machine dialog. A dialog ends when a goal is achieved by completing a transaction or by routing the user to an appropriate destination.

There are two important contributing factors to these speech technologies that enable complex services to be created in a cost-effective and consistent manner. The first is learning from data as opposed to handcrafted rules and grammars. Machine learning plays an essential role in exploiting data for generating robust and accurate systems [3]. The data can be very heterogeneous and can be originated from wizard data collections, human/human interactions, relevant Web sites, or e-mail interactions. The second important factor is the user experience (UE) design. UE can be considered as the technology front-end. It



[FIG1] Evolution of communication services.



[FIG2] Speech technologies in human/machine dialog.

plays an essential role in making the dialog progresses successfully, especially during problematic situations caused by system failures or users not providing concise or accurate information. A poor UE design can very easily overshadow the best technologies [2]. Conversely, a good UE design can make up for deficiencies in poor technologies. The combination of state-of-the-art speech and language technologies, machine learning, and good UE design results in a natural and intelligent dialog in VoiceTone applications. These applications support both call routing [4], [5], where the system understands what the callers are asking and routes them to an appropriate agent, and full transaction completion in which the system understands the caller's request and gathers all the relevant information necessary to fully automate the transaction.

ASR

The goal of ASR is to accurately and efficiently convert a speech signal into a text message, independent of the input device, speaker, or the environment. More formally, given an acoustic observation sequence X and a specified word sequence W , an optimal sequence of words can be estimated using Bayes rule:

$$\hat{W} = \arg \max_W P(W|X) = \arg \max_W P(W)P(X|W).$$

In its basic form, $P(W)$ is the language model estimated using n -gram statistics and $P(X|W)$ is the acoustic model represented by a hidden Markov model (HMM), trained using maximum likelihood estimation [6]. The features X capture primarily the spectral characteristics of the signal. They essentially include cepstrum and energy, along with their first- and second-order time derivatives. To model word pronunciation, a lexicon (or a

dictionary) is used to map sequences of phone units into words. The best word sequence \hat{W} (or word lattice) is computed using a pattern recognizer based on a standard Viterbi decoder.

For the past 50 years, ASR research has been predominantly focused on four key areas [7]:

- **Efficiency:** faster decoders, smaller memory footprint
- **Accuracy:** improved acoustic and language modeling for increased word accuracy
- **Robustness:** invariance to extraneous background noise and channel conditions as well as speaker and accent variations
- **Operational performance:** low latency, end-pointing, barge-in capability, rejection of out-of-vocabulary speech, and confidence scoring.

VoiceTone applications adopt the AT&T Watson engine, which is a speaker-independent, large-vocabulary speech recognizer [8]. The engine, which supports context-dependent, tied-state, continuous-density HMM, uses a network that is based on compositions of a set of weighted, finite-state transducers: from acoustic units into phones, from phones into words, and from words into phrases. This novel representation of the recognition network allows one to apply very efficient and general composition and determinization algorithms in both time and space [8]. Accordingly, with just a different network, the same engine can recognize anywhere from two vocabulary words to well over a million words with faster than real-time processing.

Many of the improvements in acoustic modeling and robustness over the past two decades have been led by the availability of more data and faster computations. The famous saying, "There is no data like more data," has been one the key drivers for improved speech recognition performance. State-of-the-art recognizers are now trained with over 1,000 hours of speech, compared with just a few hours of speech over a decade ago. Although the availability of mass amounts of speech data has been instrumental in improving word accuracy, the Watson engine also adopts novel algorithms for improved robustness towards channel distortion and speaker variations, as well as enhanced word discrimination and speaker and environment adaptation [10]–[12].

Although word accuracy is the most dominant measure used to evaluate speech recognizers, deploying ASR engines typically imposes several other technical challenges. These include fast decoding speed, accurate barge-in performance, low memory foot print, and accurate rejection performance of out-of-vocabulary words [13], [14]. Figure 3 shows the performance of the Watson engine (accuracy versus decoding speed) when tested on two VoiceTone applications. At an operating

point of 0.1 s CPU/audio time, the recognizer performs at 78% and 83% word accuracy for the two telecom and healthcare tasks, respectively [8].

TTS

Although automation platforms have the flexibility of playing prerecorded prompts, TTS is generally invaluable for reducing the cost of prompt generation and for speaking dynamic contents that are highly variable, such as e-mails, medical and insurance records, names, and addresses.

The goal of TTS is to render individual words intelligibly and achieve a prosody that sounds natural to the listener. The AT&T Natural Voices TTS engine [16], as well as other commercial TTS engines that have been introduced over the last few years, take a big step toward achieving naturalness through a method called *concatenative synthesis* [15]. This process includes building synthesized utterances by concatenating short snippets of recorded human speech. Concatenative synthesis, which is a departure from traditional synthesis methods such as articulatory synthesis and formant synthesis, exploits the availability of high-quality recordings as well as fast decoding algorithms to provide very high quality human-like speech.

At the highest level, a TTS engine performs two main functions: text analysis and speech synthesis. Text analysis includes text normalization, such as expanding abbreviations and eliminating nonalphabet characters, phonetic analysis for retrieving pronunciations from either a dictionary or letter-to-sound rules, and prosodic analysis for determining word accents, phrase breaks, and intonations. The speech synthesis component uses a database of small atomic units (such as phones, syllables, etc). Unit selection is accomplished by treating the units in the database as states in a fully connected transition network and then executing a Viterbi search for the path that minimizes the cost function [15].

SR

Security of information and individual identity have become the major social and business issues of the 21st century. Whether it is security of personal records, such as bank account information, or security against identity theft or within public facilities like airports or shopping malls, everyone is concerned about their safety and the safety of their information.

Human speech is an important biometric feature for actively or passively authenticating a person's identity. In many respects, speaker recognition (SR) technology has advanced to a point where it can be shown to be superior to human performance [17]. SR is a broad field of technologies that encompasses three related, but distinct, areas. Speaker verification (SV) is the process of using human voice characteristics to verify whether a person is who he or she claims to be. Speaker identification (SI) is the process of identifying a user from a closed population of several users. Speaker segmentation and detection (SSD) is the process of detecting and locating each speaker within media where more than one person may be talking.

Speaker recognition technologies can be either text dependent or text independent. In text-dependent mode, the

user is limited to saying a predefined set of known words. This is modeled using an HMM designed for each speaker [19]. In text-independent mode, the user is provided with greater flexibility to speak from an unrestricted set. Text-independent systems are modeled using a Gaussian mixture model (GMM) [18]. For both modes, the problem can be considered as a likelihood ratio test, where the null hypothesis represents a speaker model and the alternative hypothesis represents a cohort (or competitive) model.

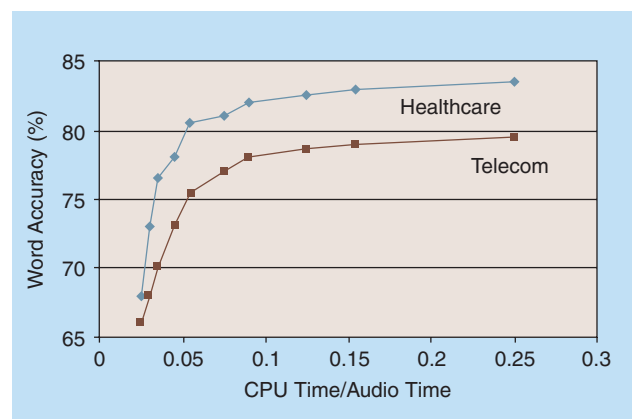
The VoiceTone platform adopts an adaptive SR technology in which initial speaker models based on limited enrollment data are robustly estimated using regularization methods. With adaptation of both speaker models and decision thresholds, SR applications are able to achieve equal error rates of less than 1% even under fairly difficult acoustic conditions [19].

SLU

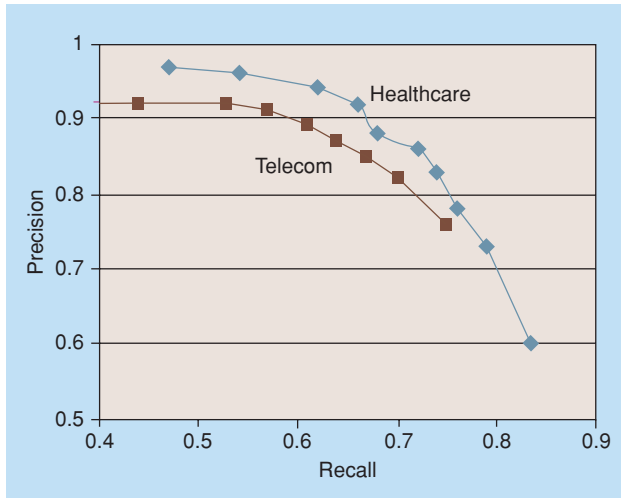
Although interest in ASR and TTS can be traced as far back as the late 1700s, when Von Kempelen created a mechanical talking machine that mimics the human voice apparatus, research in SLU only began in the early 1970s. This era saw the introduction of the Advanced Research Projects Agency (ARPA)-funded Speech Understanding Research (SUR) program. The technology became more widespread in the 1990s, when speech recognizers became powerful enough that they could operate in real-time with relatively reasonable accuracy. SLU makes it viable to adopt natural language dialog applications without having to achieve perfect recognition accuracy and without dictating what users should say, as present in traditional, system-initiative dialogs.

In SLU, the goal is to extract the meaning of the recognized speech to identify a user's request and fulfill their need. Let L be an input lattice that may combine acoustic, linguistic, semantic, and pragmatic features and C be its underlying meaning (conceptual representation). Using Bayes rule, the best conceptual structure can be defined as

$$C^* = \arg \max_C P(L|C)P(C),$$



[FIG3] ASR performance on two VoiceTone customer care applications.



[FIG4] SLU performance on two VoiceTone customer care applications.

where $P(C)$ is essentially the semantic language model and $P(L|C)$ is the lexical realization model that specifies how a lattice is generated given a meaning [20].

One of the key challenges in SLU is defining a general set of concepts (or tags) that can adequately describe the meaning space. In VoiceTone, we adopt a predicate-argument representation language that associates each lexical input with a semantic request [21]. Each predicate, or dialog act, represents a domain-independent verb, whereas each argument constitutes an object on which actions are performed. For example, “I want credit for charges on my bill” is annotated as request (BillCredit).

There are essentially three components in the VoiceTone SLU system. A preprocessor normalizes the input word lattice by removing disfluencies and applying morphological and synonym rules. An entity extractor is a sort of shallow parser that is written in a grammar rule notation. These grammars, compiled as finite-state transducers, are used to capture domain-dependent as well as domain-independent entities. For example, a named entity may be a phone number, a type of service or product, or a person’s name. These entities are detected from a preprocessed word lattice, extracted, and then normalized if necessary [21]. The last step includes transforming the output of the entity extractor into a sequence of predicate-argument labels (or semantic concepts). We consider this task as a problem of semantic classification where one can apply either generative methods or discriminative methods to identify the reason (or topic) why the user called [for example, request(balance), verify(payment), etc.].

To obtain superior generalization performance, we adopted discriminative techniques based on large margin classifiers, including boosting and support vector machines (SVM) [3]. The basic idea in boosting is to build an accurate classifier by combining “weak” or “simple” classifiers, each of which may be moderately accurate. In SVM, the goal is to select either linear or nonlinear kernels to maximize the distance (or margin) among classes in a higher dimensional space. At run time, given

an input from the entity extractor L' , the classifier outputs $P(c|L')$ for each class $c \in C$. The classifier output can be used as a confidence score to exploit different dialog strategies. Based on a desired F-measure value (trade-off between precision and recall), appropriate confidence thresholds are selected, below which the DM either confirms or reprompts the user [24].

Although machine learning systems provide added robustness and improved generality over rule-based grammars, they typically require a significant amount of training data, which is often expensive to collect and difficult to label. In VoiceTone, the semantic classifier is extended to accommodate for both application knowledge, in the form of grammar rules, and data. In boosting, for instance, the loss function is extended to include a measure of relative entropy between the probabilities predicted by the prior knowledge and those predicted by the data [28]. Figure 4 shows the semantic classification performance of the SLU engine when tested on the two customer care applications, each having about 100 semantic concepts. The performance is displayed in terms of a precision versus recall curve. At an operating point of about 90% precision, for example, the recall rate is between 60–70%. Typically, about one third of the errors made are attributed to misrecognition.

DM

ASR, TTS, SR, and SLU technologies apply to the processing of single speech utterances. The objective of a VoiceTone application, however, is to fulfill a transaction, which often requires an extended conversation with a user. A dialog manager (DM) orchestrates this conversation.

A dialog application typically follows the logic specified in a call flow designed by a human factors engineer. A popular approach for implementing directed dialog call flows consists of writing an application directly in one of the mark-up languages designed for voice communications, mainly VoiceXML [1] or speech application language tags (SALT) [22]. Several directed dialog applications on the VoiceTone platform are written in VoiceXML in this manner.

Supporting natural language dialog applications in VoiceTone, however, is more complex than the step-by-step progression typical of a directed dialog interaction; the DM is required to exploit the dialog history, support mixed-initiative interactions, and handle coreferencing. At each dialog turn, the speech recognizer engine sends a string of recognized words to the SLU engine. The SLU determines the set of semantic classes and identifies the named entities in the utterance. The output is evaluated by the AT&T DM, *Florence* [24], in the context of the call flow specification, the current state of the dialog, and possibly information from other sources (typically from an application-specific database). This results in dynamically generating a new VoiceXML page that initiates the next dialog turn.

The DM may apply a range of strategies to control the dialog flow. Generally, these strategies fall into three categories: state machines, frame- or rule-based systems, and “agent” systems that rely on reasoning or planning, each with advantages and disadvantages relative to different application tasks [25]. More recent advances in spoken dialog modeling explore new

approaches to human-machine dialog based on stochastic representation of the dialog states and reinforcement learning techniques to optimize the DM policies [20], [26].

To take advantage of multiple dialog strategies within the VoiceTone applications, the Florence DM is built around the notion of a *flow controller*. This provides an infrastructure on which multiple strategies can coexist in an application, each encapsulated in a separate flow controller and sharing common data structures. An application is built by specifying a particular call flow to one or more of these flow controllers. The primary flow controller implements an extension of a finite state machine called an *augmented recursive transition network* (ATN) [23], in which the human-machine interaction is described as a network of states and transitions. Transitions from a state are chosen based on a set of conditions (typically consisting of an interpretation of the SLU results or database contents, plus context information and dialog history). Each transition is associated with a set of dialog actions, consisting of such functions as a prompt and speech recognition, a database access, or a call transfer. Transition networks provide a compact representation of the call-routing applications and easily map to the block diagram that user-experience engineers commonly employ to specify call flows.

A second flow controller uses topic hierarchy or task ontology to represent more complex dialog strategies. It allows the system to engage in a clarification dialog with the user to resolve ambiguous requests. The flow controller execution is specified by an XML-based markup language, which is interpreted at run-time and performs interpretive processing on the SLU results in terms of its current context. It loads local variables with the appropriate call types and entities, based on thresholds, call classification hierarchies, and the current state of the dialog. It may have to resolve elliptical or anaphoric references, identify missing or irrelevant information, or prepare some part of the input for a confirmation. The history mechanism, and other functionalities common to all flow controllers, are built into a shared DM infrastructure. This infrastructure supports context shift, the ability for the application to support mixed-initiative dialog. Context shifts can occur as an error correction (“No, I do not want to pay a bill, I want a copy of the bill”) or when a user changes his/her mind midconversation (“Perhaps I should reserve a room before I book the flight . . .”).

The Florence DM is also used to build a generic, automated Wizard-of-Oz system that addresses the initial phases of an application creation process. The automated Wizard-of-Oz system is typically designed to collect human/machine interactions closely replicating field conditions. It is capable of recognizing and understanding application-independent requests by exploiting an SLU data library and providing some level of dialog repairs to effectively conduct a constructive dialog between the user and the machine. The SLU library is a hierarchical set of call types, named entities, and their respective speech data that we have collected from numerous VoiceTone applications.

EVALUATING NL SERVICES

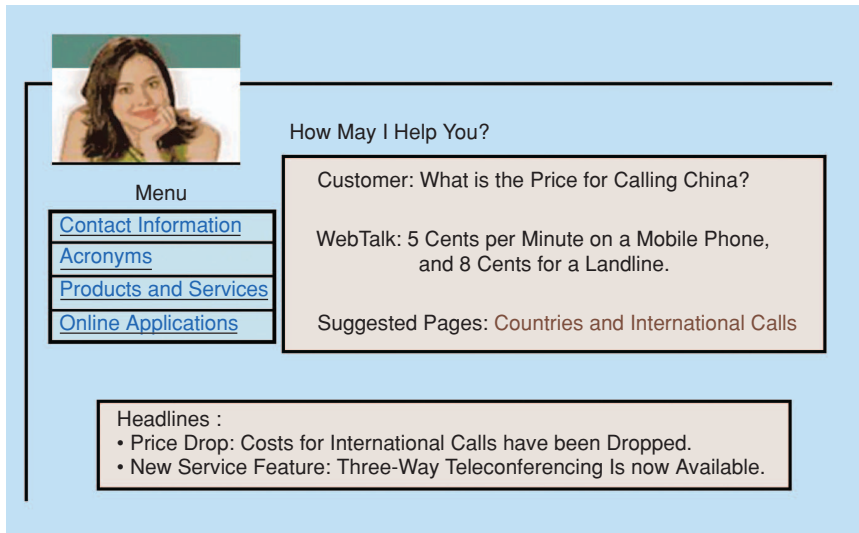
Spoken language dialog systems used as virtual agents in contact centers are resulting in significant cost savings for numer-

ous industries. Typically, call center human agents cost US\$2–\$15 per call, while automated dialog systems cost less than US\$0.20 per call and are 40% faster on average. Measures for evaluating these systems are not only realized in terms of cost reductions but also in terms of better user experiences and increased customer retention. Although the exact performances of these measures will vary as services become more mature and as users become more familiar with the new natural language capabilities, the nature of these measures will continue to evolve as we gain more experience with service deployments in various horizontal and vertical markets. In VoiceTone, for example, the integration of natural language capabilities with heterogeneous sources of information (e.g., personal account information) has been shown to result in several benefits, including increased “self-service” completion rates, fewer misdirected calls, reduced customer callbacks, and better user satisfaction. In one particular study based on AT&T customers, spoken dialog agents helped to reduce the “outpic” rate (customers who leave AT&T) by 18%. These virtual agents also reduced the average call length, which is typically around six minutes, by 10%.

FUELING THE TRENDS

Several technology trends and new technologies will fuel the growth of VoiceTone capabilities over the next few years. As the wireless network bandwidth per capita and the speed of the CPUs keep growing, rich and interactive multimedia applications will continue to migrate from the core network to the periphery hosted by mobile devices. These factors will enable ubiquitous multimedia communication services *anywhere* and *anytime* for both consumers and “road warrior” professionals. At the same time, broadband access is increasing at impressive rates and will reach 64% of American households by the end of 2005. Grid technology has also grown considerably, turning the workspace, through high-quality audio and videoconferencing, into sophisticated, room-based computing environments still affordable to large corporations and research center facilities.

Although this scenario seems to suggest that video, data, and voice services are relentlessly converging to a seamless integrated Internet Protocol (IP) network, it is still unclear what kind of applications and standards will succeed in fostering the next-generation human/machine communication services. It is certain, however, that these multimedia services, powered by a new breed of machines, will operate globally, support multiple languages, and provide input access through a variety of different modalities (such as speech, gesture, pen, etc). Automatic learning will become a key aspect of these services as they improve in performance and scale in complexity over time with minimal human intervention. Mining of heterogeneous data (call logs, speech, and call center information) will also be an integrated part of these multimedia services to help extract trends and knowledge that can aid in improving business operation. We will now address four properties of these services: 1) multimedia, multimodal, and multilingual, 2) learning, 3) automation, and 4) mining.



[FIG5] The WebTalk interactive dialog system.

MULTIMEDIA, MULTIMODAL, AND MULTILINGUAL SERVICES

Multimodal services is one of the most interesting fields that will benefit from broadband access and powerful mobile devices. Multimodal applications combine interactive multimedia output, where audio, video, graphic, and text are accurately synchronized under time constraints, and several modes of input, such as speech, keyboard, pointing devices, handwriting, and gestures. They provide richer, more versatile, and more robust user interfaces (UIs) compared to traditional unimodal speech services, offering the opportunity to adapt the UE to the actual terminal capabilities of the access device.

Creating multimodal services will pose a major technical challenge. Traditional ASR, SLU, and DM technologies would have to be revisited to factor in other inputs and technologies, such as handwriting and gesture recognition/understanding. Moreover, much research is needed in multimodal UI design, which is currently much more restrictive to multimedia and video gaming. Major players in the industry, like Microsoft and IBM, have recognized the potential of this new market, proposing limited solutions to extend the concept of Web browsing with speech input. Microsoft founded the SALT forum to promote voice applications as well as multimodal navigation in a Web browser environment. SALT interoperates with existing Web standards, introducing a few new tags for speech control. IBM and Motorola, on the other hand, proposed an alternative approach to multimodal called XHTML + Voice, or X + V. This approach extends the existing VoiceXML standards and complement them with a presentation layer built around XHTML. In 2002, W3C launched a multimodal interaction working group with the purpose of forming a common agreement upon a general architecture and an approach to define multimodal applications. The next generation of multimodal services will take advantage of integration of modalities in such a way that mutual disambiguation can lead to more robust performance.

The AT&T's Multimodal Access to City Help (MATCH) [30] framework is a multimodal dialog system test bed used to study and address most of the challenges described above, including mutual disambiguation. MATCH is a city guide and navigation system hosted on a tablet PC, which enables mobile users to access information about restaurants and subways for various states in the United States. Users can select areas of interest on a map by circling with a pen, speaking directly to the device, or through handwriting on the screen. Mutual disambiguation, when using speech and gesture inputs simultaneously, is supported using a multimodal, finite-state transducer that outputs a semantic lattice representation. This is interpreted by a multimodal DM to produce the next

dialog action. A multimodal generator renders a graphical response with eventual animation, together with a TTS synthesizer.

Besides having machines that can support multimodal and multimedia interfaces, these machines must also enable two users (e.g., an agent and a customer), speaking two different languages, to engage in a constructive dialog. AT&T Anuvaad [29], and other research systems for speech-to-speech translation, have made a significant leap in creating high-performance translation machines that can learn from mass amounts of parallel corpora. Anuvaad, for example, employs a unified stochastic finite-state transducer that integrates the ASR language model with the translation model of a speech translator. Creating integrated services that support multimodal, multimedia, and multilingual capabilities is the main focus of the Global Autonomous Language Exploitation (GALE) program sponsored by the Defense Advanced Research Projects Agency (DARPA).

AUTOMATION

The task of creating customized spoken dialog applications is traditionally labor intensive, requiring significant data resources and tremendous levels of expertise. As a result, despite efforts to modularize and reuse components of the dialog, it is not surprising that only a few hundred speech services were actually deployed in 2004 for large business customers. These services tend to be highly customized and are typically designed independently of any other sources of information, such as a Web site or human/human conversational data.

AT&T has been pursuing a new research direction to scale this industry, completely automating the process of creating spoken, or chat-based, dialog applications by leveraging the wealth of information on business Web sites. Given that most businesses maintain a Web site, the goal has been to leverage information to create and maintain a new line of automated services that require no human intervention. A prototype system incorporating these new types of services, WebTalk [31], is shown in Figure 5. WebTalk employs speech and language technology as well as

machine learning to engage with users in a chat-based or a spoken-language dialog. It includes a Web site analyzer that can automatically download Web pages and construct dialog-oriented task knowledge. It exploits the content and structure of the Web site to generate structured and semistructured task data. A Web-site parser is applied to each Web page to identify and extract information units. Those units are broad categories of visual blocks, such as menus, questions/answers, and topic definitions.

Information extraction is a key component of the WebTalk system. It enables the system to extract key entities, such as the list of products and services, contact information, abbreviations, and frequently asked questions. This information, along with a general-purpose DM, enables WebTalk to exchange in a natural language dialog without actually involving a person in the loop either during the creation or the maintenance process.

THE ACTIVE LEARNING FAMILY

The availability of large amounts of data is a key driver in the timely creation of cost-effective, natural-language applications in VoiceTone. This data, whether in audio or text form, not only is an essential source for improving the underlying speech and language models, but it is also critical for identifying problematic dialog areas that may require immediate attention. Logging, transcribing, storing, and managing this data, however, are often difficult tasks when scaling natural language dialog applications.

As the technologies continue to become more statistical, relying more heavily on speech data as opposed to knowledge rules, the quality and the transcription accuracy of this data will continue to play a central role in the success of deployed services. In addition, speed to market will continue to be essential not only for scaling these applications, but also for providing product differentiation.

In VoiceTone services, we apply a general learning framework on the daily feed of dialog data to enable us to perform three operations (see Figure 6):

- 1) *Active learning*: Minimize the labor effort by identifying only interesting or problematic dialogs, which when manually transcribed and labeled, would result in a significant improvement in system performance. This is applied on the live daily feed of the data to help prioritize the necessary labeling effort [27].
- 2) *Active labeling*: Identify inconsistencies (or noise) in the manually labeled data. This process is applied for detecting outliers and incorrectly labeled data.
- 3) *Active evaluation*: Generating automatic alerts to indicate when an adaptive system, continuously training through active and unsupervised learning, is able to outperform the deployed production system.

This learning family of methods can be formulated in terms of cost and reward functions, as typically done in any optimization problem. Given a corpus of dialogs D , the objective is to extract a subset \bar{D} that, when transcribed and labeled, would provide a maximum improvement in performance. One may formally define a reward function $R(D)$ that includes a cost $C(D)$ for manually labeling D , and a resulting performance gain $G(D)$. $C(D)$ can be estimated in terms of the cost of the labeling effort (e.g., dollar spent per hour of labeling), while $G(D)$ can be computed from some metric of dialog success (e.g., dialog-level confidence score, which is typically synchronized with system performance). The subset of dialogs that are to be labeled is identified such that

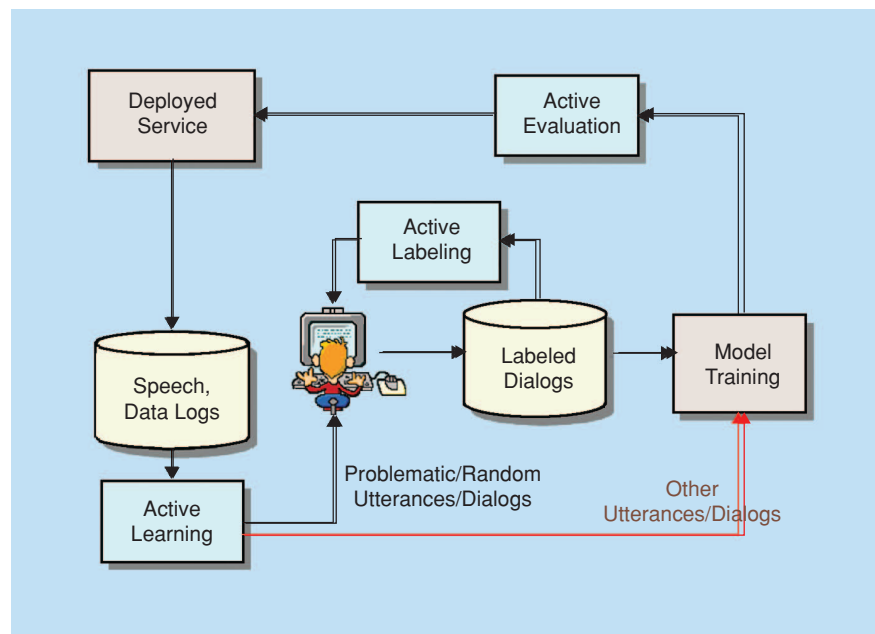
$$\bar{D} = \underset{D}{\operatorname{argmax}} R(D).$$

SPEECH MINING

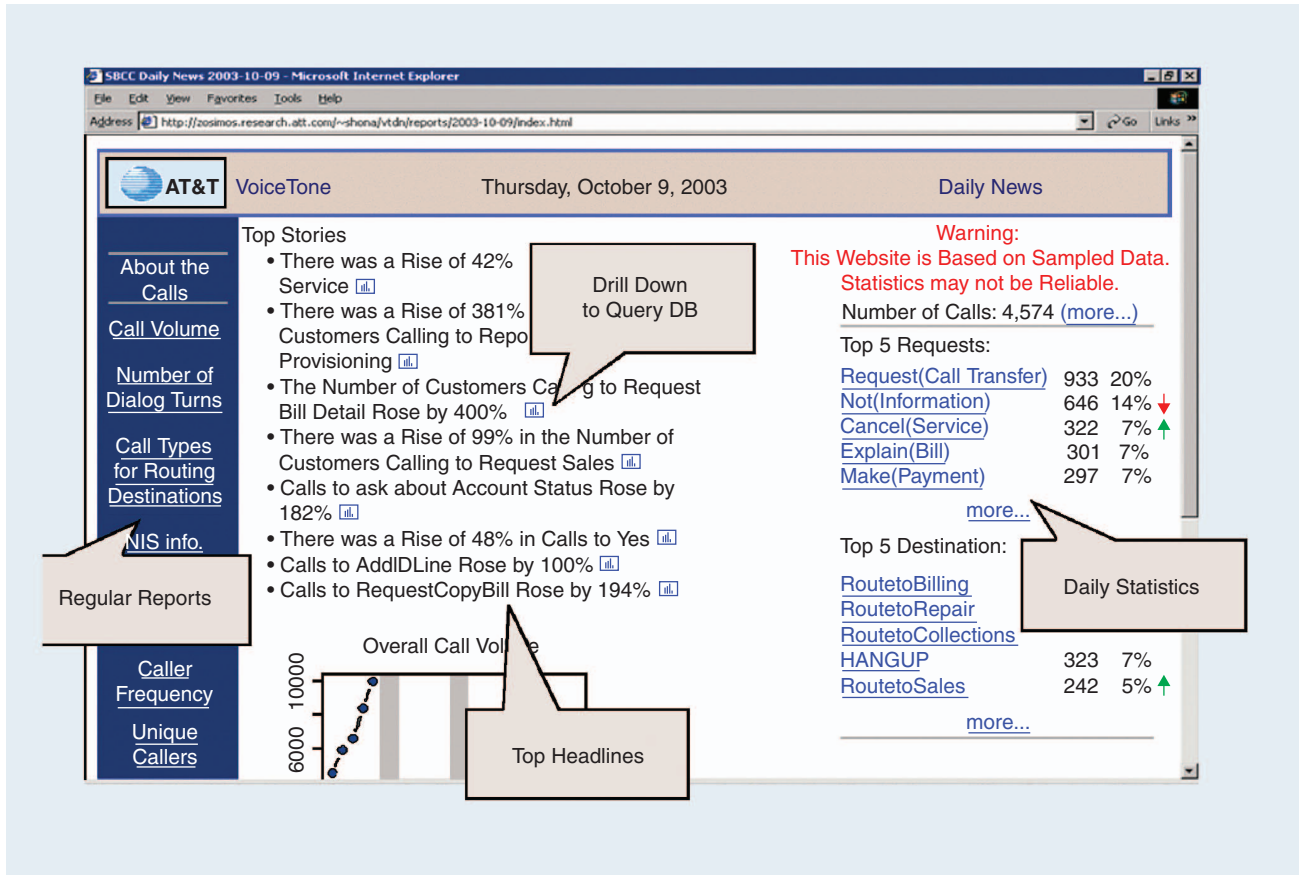
As the volume of accumulated data in contact centers grows, we are faced with the analytical challenge of finding information of interest to the researchers, developers, executives, and customers that can explain what is happening and why. Mining heterogeneous spoken dialog data for the purpose of improving system operation and extracting business intelligence is a new opportunity for conducting data mining research [34].

In AT&T VoiceTone, we have created an information mining visualization tool known as the Daily News (DN) [32]. This Web-based tool exploits spoken dialog interaction logs along with meta information, including customer profile, to automatically detect interesting and unexpected patterns or to automatically predict the onset of change. These factors are presented in a daily Web-based newsletter intended to resemble online news.

An example of the DN is shown in Figure 7. The “Top Stories,” displayed in the middle of the page, are headlines (or alerts) that are automatically generated to reflect interesting or



[FIG6] The active learning family of methods.



[FIG7] The Daily News information mining tool.

unpredictable changes in the data streams. Alerts are generated if the most recent prevalence of a feature/value pair lies outside the expected interval computed over a historic period. Each dialog is encoded by a set of features that includes the number of turns in the dialog, the length of the call and of each turn, the final disposition of the call, prompt identification, ASR and SLU result, and so on. Bivariate analysis is performed on these features using statistical process control techniques. Also on the page are tables of measurements that are of ongoing interest, such as the top five requests made by callers or the top five call-routing destinations. The headlines and other items on the page are linked to plots that show the trend over time of the feature being described, as well as the bounds of the “normal” range that are used to trigger a change condition. It is also possible to drill down on any of the items to retrieve more detailed information, including a display of the relevant individual dialogs and plots to compare user-selected conditions.

One aspect of the drill-down capability is performing a dialog trajectory analysis (DTA) [33]. DTA provides a visual representation for a set of dialogs in terms of state sequences and user responses. Call trajectories are composed and minimized using stochastic finite state automata [8], with the nodes representing system states and arcs representing user responses. Each arc is annotated with its frequency. To reduce the complexity of the displayed network, the graphical representation is

designed in a layered approach, starting with individual subdialogs down to an individual dialog turn or audio file. Arcs and nodes also highlight abnormal local dialog changes, helping to visualize the root cause of problems.

CONCLUSIONS

The explosion of multimedia data, the continuous growth in computing power, and advances in machine learning and speech and natural language processing are making it possible to create a new breed of virtual intelligent agents capable of performing sophisticated and complex tasks that are radically transforming contact centers. These virtual agents are enabling ubiquitous and personalized access to communication services from anywhere. They will ultimately provide a vehicle to fully automate eContact services without agent personnel. They will not be limited to multimodal, multimedia, and multilingual capabilities, but will also possess learning and data-mining capabilities to enable them to scale and self-maintain as well as extract and report on business intelligence. AT&T VoiceTone is a subset of this eContact revolution focused on creating this new wave of intelligent communication services.

ACKNOWLEDGMENTS

The authors would like to acknowledge the numerous technical contributions by our AT&T VoiceTone Research and

Development colleagues. We would also like to thank the associate editors for their help and guidance.

AUTHORS

Mazin Gilbert (formerly Mazin Rahim) received the Ph.D. degree in electrical engineering and electronics from the University of Liverpool, U.K. He is the director of the Spoken Language Understanding and Dialog Division at AT&T Labs Research. He has over 80 technical publications and is the author of the book *Artificial Neural Networks for Speech Analysis/Synthesis* (London: Chapman and Hall, 1994). He holds 11 U.S. patents and has received several national and international awards. He is a Senior Member of the IEEE. He is chair of the CAIP Industry Board at Rutgers University, chair of the IEEE Speech Technical Committee, and a teaching professor at Princeton University.

Jay G. Wilpon is director of the Speech Services Research Department within AT&T Laboratories. He has over 90 publications and is the coeditor of the book *Voice Communication Between Humans and Machines* (NAS Press). He received the IEEE 1987 ASSP Society's Paper Award and was chair of the IEEE Signal Processing Society's (SPS) Speech Processing Technical Committee, 1993–1995. He was member-at-large to the IEEE SPS Board of Governors from 1996 to 1999. He is an IEEE Fellow. In 2002, he was awarded the AT&T Science and Technology Medal for technical contributions and leadership in introducing automatic speech recognition services into network telecommunications.

Benjamin Stern graduated in physics (Ph.D.) from Columbia University in 1988. From 1998 to the present, he has worked in AT&T Labs. He is currently a member of the Speech Services Research Laboratory in Florham Park, NJ. He has worked on new speech-enabled communications service concepts, and on commercializing speech services within AT&T.

Giuseppe Di Fabbrizio graduated in electrical engineering (M.S.E.E.) from the Politecnico di Torino, Italy in 1990. From 1990 to 1995, he was a senior researcher with Telecom Lab Italia (formerly CSELT), Torino. In January 1996, he joined AT&T Labs-Research in Florham Park, NJ, as senior researcher where he is currently member of the Speech Services Research Laboratory. During his career he mainly conducted research on spoken dialogue systems, natural language processing and speech services, and published more than 30 papers on these subjects. He is a member the Association for Computational Linguistics (ACL) and a Senior Member of the IEEE.

REFERENCES

- [1] S. McGlashan, Voice Extensible Markup Language (VoiceXML) Version 2.0 [Online]. Available: <http://www.w3.org/TR/2004/PR-voicexml20-20040203>
- [2] M.H. Cohen, J.P. Giangola, and J. Balogh, *Voice User Interface Design*. Boston: Addison Wesley, 2004.
- [3] A. Smola, P. Bartlett, B. Scholkopf, and D. Schuurmans, *Advances in Large Margin Classifiers*. Cambridge, MA: MIT Press, 2000.
- [4] A. Gorin, G. Riccardi, and J. Wright, "How may I help you?," *Speech Commun.*, vol. 23, pp. 113–127, 1997.
- [5] Natarajan, R. Prasad, B. Suhm, and D. McCarthy, "Speech enabled natural language call routing: {BBN} call director," in *Proc. Int. Conf. Spoken Language Processing*, Denver, 2002.
- [6] L. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*. Englewood Cliffs, NJ: Prentice-Hall, 1993.

- [7] X.D. Huang, A. Acero, and H.-W. Hon, *Spoken Language Processing*. Englewood Cliffs, NJ: Prentice-Hall, 2001.
- [8] F.C. Pereira and M. Riley, "Speech recognition by composition of weighted finite automata," in *Finite-State Devices for Natural Language Processing*, Roche, E. and Schabes, Y., Eds. Cambridge, MA: MIT Press, 1997.
- [9] V. Goffin, C. Allauzen, E. Bocchieri, D. Hakkani-Tur, A. Ljolje, S. Parthasarathy, M. Rahim, G. Riccardi, and M. Saraclar, "The AT&T Watson speech recognizer," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, Philadelphia, PA, May 2005.
- [10] P.C. Woodland and D. Povey, "Large scale MMIE training for conversational telephone speech recognition," in *Proc. Speech Transcription Workshop*, 2000.
- [11] M. Rahim and B.-H. Juang, "Signal bias removal by maximum likelihood estimation for robust speech recognition," *IEEE Trans. Speech Audio Processing*, vol. 4, no. 1, 1996.
- [12] L. Lee and R. Rose, "A frequency warping approach to speaker normalization," *IEEE Trans. Speech Audio Processing*, vol. 6, pp. 49–60, Jan. 1998.
- [13] E.L. Bocchieri, "Vector quantization for the efficient computation of continuous density likelihoods," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, Apr. 1993, pp. 692–695.
- [14] S. Furui, "Recent advances in spontaneous speech recognition and understanding," in *Proc. ISCA-IEEE Workshop Spontaneous Speech Processing and Recognition*, Tokyo, Japan, 2003.
- [15] A. Hunt and A. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, Atlanta, GA, 1996, vol. 1, pp. 373–376.
- [16] M. Beutnagel, A. Conkie, J. Schroeter, Y. Stylianou, and A. Syrdal, "The AT&T next-gen TTS system," in *Proc. Joint Meeting ASA, EAA, and DEGA*, Berlin, Germany, Mar. 1999, pp. 18–24.
- [17] A. Schmidt-Nielsen and T.H. Crystal, "Speaker verification by human listeners: Experiments comparing human and machine performance using the NIST 1998 speaker evaluation data," *Digital Signal Processing*, pp. 249–266, 2000.
- [18] D.A. Reynolds and R.C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," *IEEE Trans. Speech Audio Processing*, vol. 3, no. 1, pp. 72–83, Jan. 1995.
- [19] S. Parthasarathy and A.E. Rosenberg, "General phrase speaker verification using subword background models and likelihood-ratio scoring," in *Proc. Int. Conf. Spoken Language Processing*, Philadelphia, PA, 1996, pp. 2403–2406.
- [20] S. Young, "Talking to machines (statistically speaking)," in *Proc. Int. Conf. Spoken Language Processing*, 2002.
- [21] N. Gupta, G. Tur, D. Hakkani-Tür, S. Bangalore, G. Riccardi, and M. Rahim, "The AT&T spoken language understanding system," *Accepted in IEEE Trans. Speech and Audio Processing*, 2005.
- [22] Speech Application Language Tags (SALT) 1.0 Specification [Online]. Available: <http://www.saltforum.org/saltforum/downloads/SALT1.0.pdf>
- [23] D. Bobrow and B. Fraser, "An augmented state transition network analysis procedure," in *Proc. IJCAI*, Washington, DC, May 1969, pp. 557–567.
- [24] G. Di Fabbrizio and C. Lewis, "Florence: A dialogue manager framework for spoken dialogue systems," in *Proc. 8th Int. Conf. Spoken Language Processing*, Korea, Oct. 4–8, 2004.
- [25] M.F. McTear, "Spoken dialogue technology: Enabling the conversational user interface," *ACM Comput. Surveys*, vol. 34, pp. 90–169, 2002.
- [26] E. Levin, R. Pieraccini, and W. Eckert, "A stochastic model of human-machine interaction for learning dialog strategies," *IEEE Trans. Speech Audio Processing*, vol. 8, no. 1, pp. 11–23, Jan. 2000.
- [27] G. Riccardi and D. Hakkani-Tür, "Active and unsupervised learning for automatic speech recognition," in *Proc. 8th European Conf. Speech Communication and Technology*, Geneva, Switzerland, Sept. 2003.
- [28] R. Schapire, M. Rochery, M. Rahim, and N. Gupta, "Incorporating prior knowledge into boosting," in *Proc. 19th Int. Conf. Machine Learning*, 2002.
- [29] S. Bangalore and G. Riccardi, "Stochastic finite-state models for spoken language machine translation," *Machine Translat.*, vol. 17, no. 3, pp. 165–184, 2002.
- [30] M. Johnston, S. Bangalore, G. Vasireddy, A. Stent, P. Ehlen, M. Walker, S. Whittaker, and P. Maloor, "MATCH: An architecture for multimodal dialogue systems," in *Proc. 40th Ann. Meeting Association for Computational Linguistics*, 2002, pp. 376–383.
- [31] J. Feng, S. Bangalore, and M. Rahim, "WebTalk: Towards automatically building dialog services by exploiting the content and structure of websites," in *Proc. Int. Conf. World Wide Web*, Budapest, Hungary, May 2003.
- [32] S. Douglas, D. Agarwal, T. Alonso, R. Bell, M. Rahim, D.F. Swayne, and C. Volinsky, "Mining customer care dialogs for Daily News," in *Proc. Int. Conf. Spoken Language Processing*, Korea, 2004.
- [33] A. Abella, J. Wright, and A. Gorin, "Dialog trajectory analysis," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, Canada, 2004.
- [34] I. Witten and E. Frank, *Data Mining*, San Mateo, CA: Morgan Kaufmann, 1999.