SALT: An XML Application for Web-based Multimodal Dialog Management

Kuansan Wang
Speech Technology Group, Microsoft Research
One Microsoft Way, Microsoft Corporation
Redmond, WA, 98006, USA
http://research.microsoft.com/stg

Abstract

This paper describes the Speech Application Language Tags, or SALT, an XML based spoken dialog standard for multimodal or speech-only applications. A key premise in SALT design is that speech-enabled user interface shares a lot of the design principles and computational requirements with the graphical user interface (GUI). As a result, it is logical to introduce into speech the object-oriented, event-driven model that is known to be flexible and powerful enough in meeting the requirements for realizing sophisticated GUIs. By reusing this rich infrastructure, dialog designers are relieved from having to develop the underlying computing infrastructure and can focus more on the core user interface design issues than on the computer and software engineering details. The paper focuses the discussion on the Web-based distributed computing environment and elaborates how SALT can be used to implement multimodal dialog systems. How advanced dialog effects (e.g., cross-modality reference resolution, implicit confirmation, multimedia synchronization) can be realized in SALT is also discussed.

Introduction

Multimodal interface allows a human user to interaction with the computer using more than one input methods. GUI, for example, is multimodal because a user can interact with the computer using keyboard, stylus, or pointing devices. GUI is an immensely successful concept, notably demonstrated by the World Wide Web. Although the relevant technologies for the Internet had long existed, it was not until the adoption of GUI for the Web did we witness a surge on its usage and rapid improvements in Web applications.

GUI applications have to address the issues commonly encountered in a goal-oriented dialog system. In other words, GUI applications can be viewed as conducting a dialog with its user in an iconic language. For example, it is very common for an application and its human user to undergo many exchanges before a task is completed. The application therefore must manage the interaction history in order to properly infer user's intention. The interaction style is mostly system initiative because the user often has to follow the prescribed interaction flow where allowable branches are visualized in graphical icons. Many applications have introduced mixed initiative features such as type-in help or search box. However, user-initiated digressions are often recoverable only if they are anticipated by the application designers. The plan-based dialog theory (Sadek *et al* 1997, Allen 1995, Cohen *et al* 1989) suggests that, in order for the mixed initiative dialog to function properly, the computer and the user should be collaborating partners that actively assist each other in planning the dialog flow. An application will be perceived as hard to use if the flow logic is obscure or unnatural to the user and, similarly, the user will feel frustrated if the methods to express intents are too limited. It is widely believed that spoken language can improve the user interface as it provides the user a natural and less restrictive way to express intents and receive feedbacks.

The Speech Application Language Tags (SALT 2002) is a proposed standard for implementing spoken language interfaces. The core of SALT is a collection of objects that enable a software program to listen, speak, and communicate with other components residing on the underlying platform (e.g., discourse manager, other input modalities, telephone interface, etc.). Like their predecessors in the Microsoft Speech Application Interface (SAPI), SALT objects are programming language independent. As a result, SALT objects can be embedded into a HTML or any XML document as the spoken language interface (Wang 2000). Introducing speech capabilities to the Web is not new (Aron 1991, Ly *et al* 1993, Lau *et al* 1997). However, it is the utmost design goal of SALT that advanced dialog management techniques (Sneff *et al* 1998, Rudnicky *et al* 1999, Lin *et al* 1999, Wang 1998) can be realized in a straightforward manner in SALT.

The rest of the paper is organized as follows. In Sec. 1, we first review the dialog architecture on which the SALT design is based. It is argued that advanced spoken dialog models can be realized using the Web infrastructure. Specifically, various stages of dialog goals can be modeled as Web pages that the user will navigate through. Considerations in flexible dialog designs have direct implications on the XML document structures. How SALT implements these document structures are outlined. In Sec. 2, the XML objects providing spoken language understanding and speech synthesis are described. These objects are designed using the event driven architecture so that they can included in the GUI environment for multimodal interactions. Finally in Sec. 3, we describe how SALT, which is based on XML, utilizes the extensibility of XML to allow new extensions without losing document portability.

1 Dialog Architecture Overview

With the advent of XML Web services, the Web has quickly evolved into a gigantic distributed computer where Web services, communicating in XML, play the role of reusable software components. Using the universal description, discovery, and integration (UDDI) standard, Web services can be discovered and linked up dynamically to collaborate on a task. In other words, Web services can be regarded as the software agents envisioned in the open agent architecture (Bradshaw 1996). Conceptually, the Web infrastructure provides a straightforward means to realize the agent-based approach suitable for modeling highly sophisticated dialog (Sadek *et al* 1997). This distributed model shares the same basis as the SALT dialog management architecture.

1.1 Page-based Dialog Management

An examination on human to human conversation on trip planning shows that experienced agents often guide the customers in dividing the trip planning into a series of more manageable and relatively untangled subtasks (Rudnicky *et al* 1999). Not only the observation contributes to the formation of the plan-based dialog theory, but the same principle is also widely adopted in designing GUI-based transactions where the subtasks are usually encapsulated in visual pages. Take a travel planning Web site for example. The first page usually gathers some basic information of the trip, such as the traveling dates and the originating and destination cities, etc. All the possible travel plans are typically shown in another page, in which the user can negotiate on items such as the price, departure and arrival times, etc. To some extent, the user can alter the flow of interaction. If the user is more flexible for the flight than the hotel reservation, a well designed site will allow the user to digress and settle the hotel reservation before booking the flight. Necessary confirmations are usually conducted in separate pages before the transaction is executed.

The designers of SALT believe that spoken dialog can be modeled by the page-based interaction as well, with each page designed to achieve a sub-goal of the task. There seems no reason why the planning of the dialog cannot utilize the same mechanism that dynamically synthesizes the Web pages today.

1.2 Separation of Data and Presentation

SALT preserves the tremendous amount of flexibility of a page-based dialog system in

dynamically adapting the style and presentation of a dialog (Wang 2000). A SALT page is composed of three portions: (1) a data section corresponding to the information the system needs to acquire from the user in order to achieve the sub-goal of the page; (2) a presentation section that, in addition to GUI objects, contains the templates to generate speech prompts and the rules to recognize and parse user's utterances; (3) a script section that includes inference logic for deriving the dialog flow in achieving the goal of the page. The script section also implements the necessary procedures to manipulate the presentation sections.

This document structure is motivated by the following considerations. First, the separation of the presentation from the rest localizes the natural language dependencies. An application can be ported to another language by changing only the presentation section without affecting other sections. Also, a good dialog must dynamically strike a balance between system initiative and user initiative styles. However, the needs to switch the interaction style do not necessitate changes in the dialog planning. The SALT document structure maintains this type of independence by separating the data section from the rest of the document, so that when there are needs to change the interaction style, the script and the presentation sections can be modified without affecting the data section. The same mechanism also enables the app to switch among various UI modes, such as in the mobile environments where the interactions must be able to seamlessly switching between a GUI and speech-only modes for hand-eye busy situations. The presentation section may vary significantly among the UI modes, but the rest of the document can remain largely intact.

1.3 Semantic Driven Multimodal Integration

SALT follows the common GUI practice and employs an object-oriented, event-driven model to integrate multiple input methods. The technique tracks user's actions and reports them as events. An object is instantiated for each event to describe the causes. For example, when a user clicks on a graphical icon, a mouse click event is fired. The mouse-click event object contains information such as coordinates where the click takes place. SALT extends the mechanism for speech input, in which the notion of semantic objects (Wang 2000, Wang 1998) is introduced to capture the meaning of spoken language. When the user says something, speech events, furnished with the corresponding semantic objects, are reported. The semantic objects are structured and categorized. For example, an utterance "Send mail to John" is composed of two nested semantic objects: "John" representing the semantic type "Person" and the whole utterance the semantic type "Email command." SALT therefore enables a multimodal integration algorithm based on semantic type compatibility (Wang 2001). The same command can be manifest in a multimodal expression, as in "Send email to him [click]" where the email recipient is given by a point-and-click gesture. Here the semantic type provides a straightforward way to resolve the cross modality reference: the handler for the GUI mouse click event can be programmed into producing a semantic object of the type "Person" which can subsequently be identified as a constituent of the "email command" semantic object. Because the notion of semantic objects is quite generic, dialog designers should find little difficulty employing other multimodal integration algorithms, such as the unification based approach described in (Johnston et al 1997), in SALT.

2 Basic Speech Elements in SALT

SALT speech objects encapsulate speech functionality. They resemble to the GUI objects in many ways. Because they share the same high level abstraction, SALT speech objects interoperate with GUI objects in a seamless and consistent manner. Multimodal dialog designers can elect to ignore the modality of communication, much the same way as they are insulated from having to distinguish whether a text string is entered to a field through a keyboard or cut and pasted with a pointing device.

2.1 The Listen Object

The "listen" object in SALT is the speech input object. The object must be initialized with a speech grammar that defines the language model

and the lexicon relevant to the recognition task. The object has a *start* method that, upon invocation, collects the acoustic samples and performs speech recognition. If the language model is a probabilistic context free grammar (PCFG), the object can return the parse tree of the recognized outcome. Optionally, dialog designers can embed XSLT templates or scripts in the grammar to shape the parse tree into any desired format. The most common usage is to transform the parse tree into a semantic tree composed of semantic objects.

A SALT object is instantiated in an XML document whenever a tag bearing the object name is encountered. For example, a listen object can be instantiated as follows:

```
<listen id="foo" onreco="f()"
        onnoreco="g()" mode="automatic">
    <grammar src="../meeting.xml"/>
</listen>
```

The object, named "foo," is given a speech grammar whose universal resource indicator (URI) is specified via a <grammar> constituent. As in the case of HTML, methods of an object are invoked via the object name. For example, the command to start the recognition is foo.start() in the ECMAScript syntax. Upon a successful recognition and parsing, the listen object raises the event "onreco." The event handler, f(), is associated in the HTML syntax as shown above. If the recognition result is rejected, the listen object raises the "onnoreco" event, which, in the above example, invokes function g(). As mentioned in Sec. 0, these event handlers reside in the script section of a SALT page that manages the within-page dialog flow. Note that SALT is designed to be agnostic to the syntax of the eventing mechanism. Although the examples through out this article use HTML syntax, SALT can operate with other eventing standards, such as World Wide Web Consortium (W3C) XML Document Object Model (DOM) Level 2, ECMA Common Language Infrastructure (CLI), or the upcoming W3C proposal called XML Events.

The SALT listen object can operate in one of the three modes designed to meet different UI requirements. The *automatic* mode, shown above, automatically detects the end of utterance and cut off the audio stream. The mode is most suitable for push-to-talk UI or telephony based systems. Reciprocal to the *start* method, the listen object also has a *stop* method for forcing the recognizer to stop listening. The designer can explicitly invoke the stop method and not rely on the recognizer's default behavior. Invoking the stop method becomes necessary when the listen object operates under the *single* mode, where the recognizer is mandated to continue listening until the stop method is called. Under the single mode, the recognizer is required to evaluate and return hypotheses based on the full length of the audio, even though some search paths may have reached a legitimate end of sentence token in the middle of the audio stream. In contrast, the third *multiple* mode allows the listen object to report hypotheses as soon as it sees fit. The single mode is designed for push-hold-and-talk type of UI, while the multiple mode is for real-time or dictation type of applications.

The listen object also has methods to modify the PCFG it contains. Rules can be dynamically activated and deactivated to control the perplexity of the language model. The semantic parsing templates in the grammar can be manipulated to perform simple reference resolution. For example, the grammar below (in SAPI format) demonstrates how a deictic reference can be resolved inside the SALT listen object:

```
<rule propname="drink" …>
  <option> the </option>
  <list>
     <phrase propvalue="coffee"> left </phrase>
     <phrase propvalue="juice"> right </phrase>
  </list>
  <option> one </option>
</rule>
```

In this example, the *propname* and *propvalue* attributes are used to generate the semantic objects. If the user says "the left one," the above grammar directs the listen object to return the semantic object as <drink text="the left one">coffee</drink>. This mechanism for composing semantic objects is particularly useful for processing expressions closely tied to how

data are presented. The grammar above may be used when the computer asks the user for choice of the drink by displaying the pictures of the choices side by side. However, if the display is tiny, the choices may be rendered as a list, to which a user may say "the first one" or "the bottom one." SALT allows dialog designers to approach this problem by dynamically adjusting the speech grammar.

2.2 The prompt object

The SALT "prompt" object is the speech output object. Like the listen object, the prompt object has a *start* method to begin the audio playback. The prompt object can perform text to speech synthesis (TTS) or play pre-recorded audio. For TTS, the prosody and other dialog effects can be controlled by marking up the text with synthesis directives.

Barge-in and bookmark are two events of the prompt object particularly useful for dialog designs. The prompt object raises a barge-in event when the computer detects user utterance during a prompt playback. SALT provides a rich program interface for the dialog designers to specify the appropriate behaviors when the barge-in occurs. Designers can choose whether to delegate SALT to cut off the outgoing audio stream as soon as speech is detected. Delegated cut-off minimizes the barge-in response time, and is close to the expected behavior for users who wish to expedite the progress of the dialog without waiting for the prompt to end. Similarly, non-delegated barge-in let the user change playback parameters without interrupting the output. For example, the user can adjust the speed and volume using speech commands while the audio playback is in progress. SALT will automatically turn on echo cancellation for this case so that the playback has minimal impacts on the recognition.

The timing of certain user action or the lack thereof often bears semantic implications. Implicit confirmation is a good example, where the absence of an explicit correction from the user is considered as a confirmation. The prompt object introduces an event for reporting the landmarks of the playback. The typical way of catching the playback landmarks in SALT is as such:

```
<prompt id="bar" onbookmark="f()" …>
   Traveling to New York?
   <bookmark name="imp_confirm"/>
   There are <emph> 3 </emph> flights available
   …
</prompt>
```

When the synthesizer reaches the TTS markup <bookmark>, the onbookmark event is raised and the event hander f() is invoked. When a barge-in is detected, the dialog designer can determine if the barge-in occurs before or after the bookmark by inspecting whether the function f() has been called or not.

Multimedia synchronization is another main usage for TTS bookmarks. When the speech output is accompanied with, for example, graphical animations, TTS bookmarks are an effective mechanism to synchronize these parallel outputs.

To include dynamic content in the prompt, SALT adopts a simple template-based approach for prompt generation. In other words, the carrier phrases can be either pre-recorded or hard-coded, while the key phrases can be inserted and synthesized dynamically. The prompt object that confirms a travel plan may appear as the following in HTML:

```
<input name="origin" type="text" />
<input name="destination" type="text" />
<input name="date" type="text" />
…
<prompt …> Do you want to fly from
   <value targetElement="origin"/> to
   <value targetElement="destination"/> on
   <value targetElement="date"/>?
</prompt>
```

As shown above, SALT uses a <value> tag inside a prompt object to refer to the data contained in other parts of the SALT page. In this example, the prompt object will insert the values in the HTML input objects in synthesizing the prompt.

2.3 Declarative Rule-based Programming

Although the examples use procedural programming in managing the dialog flow control, SALT designers can practice inference programming in a declarative rule-based fashion in which rules are attached to the SALT objects capturing user's actions, e.g., the listen object. Instead of authoring procedural event handlers, designers can declare inside the listen object rules that will be evaluated and invoked when the semantic objects are returned. This is achieved through a SALT <bind> element as demonstrated below:

```
<listen …> <grammar …/>
   <bind test="/@confidence $lt$ 50"
         targetElement="prompt_confirm"
         targetMethod="start"
         targetElement="listen_confirm"
         targetMethod="start" />
   <bind test="/@confidence $ge$ 50"
         targetElement="origin"
         value="/city/origin"
         targetElement="destination"
         value="/city/destination"
         targetElement="date"
         value="/date" />     …
</listen>
```

The predicate of each rule is applied in turns against the result of the listen object. They are expressed in the standard XML Pattern language in the "test" clause of the <bind> element. In this example, the first rule checks if the confidence level is above the threshold. If not, the rule activates a prompt object (*prompt_confirm*) for explicit confirmation, followed by a listen object *listen_confirm* to capture the user's response. The speech objects are activated via the *start* method of the respective object. Object activations are specified in the *targetElement* and the *targetMethod* clauses of the <bind> element. Similarly, the second rule applies when the confidence score exceeds the prescribed level. The rule extracts the relevant semantic objects from the parsed outcome and assigns them to the respective elements in the SALT page. As shown above, SALT reuses the W3C XPATH language for extracting partial semantic objects from the parsed outcome.

3 SALT Extensibilities

Naturally spoken language is a modality that can be used in widely diverse environments where user interface constraints and capabilities vary significantly. As a result, it is only practical to define into SALT the speech functions that are universally applicable and implementable. For example, the basic speech input function in SALT only deals with speaker independent recognition and understanding, even though speaker dependent recognition or speaker verification are in many cases very useful. As a result, extensibility is crucial to a natural language interface like SALT.

SALT follows the XML standards that allow extensions being introduced on demand without sacrificing document portability. Functions that are not already defined in SALT can be introduced at the component level, or as a new feature of the markup language. In addition, SALT requires the standard of XML be followed so that extensions can be identified and the methods to process the extensions can be discovered and integrated.

3.1 Component extensibility

SALT components can be extended with new functions individually through the component configuration mechanism of the <param> element. For example, the <listen> element has an event to signal when speech is detected in the incoming audio stream. However, the standard does not specify an algorithm for detecting speech. A SALT document author, however, can declare reference cues so that the document can be rendered in the similar way among different processors. The <param> element can be used to set the reference algorithm and threshold for detecting speech in the <listen> object:

```
<listen onspeechdetected = "handler( )" …>
   <param xmlns:xyz="urn://xyz.edu/algo-1">
         <xyz:method>energy</xyz:method>
         <xyz:threshold>0.7</xyz:threshold>
   </param>  …
</listen>
```

Here the parameters are set using an algorithm whose uniform resource name (URN),

xyz.edu/algo-1, is declared as an attribute of the XML namespace of <param>. The parameters for configuring this specific speech detection method are further specified in the child elements. A document processor can perform a schema translation on the URN namespace into any schema the processor understands. For example, if the processor implements the speech detection algorithm where the detection threshold has a different range, the adjustment can be easily made when the document is parsed.

The same mechanism is used to extend the functionality. For instance, the <listen> object can be used for speaker verification because the algorithm used for verification and the programmatic interfaces share a lot in common with the recognition. In SALT, a <listen> object can be extended for speaker verification through configuration parameters:

```
<listen  onreco="success( )" onerror="na( )"
         onnoreco="failed( )">
    <param xmlns:v="urn:abc.com/spkrveri">
         <v:cohort>../../data</v:cohort>
         <v:threshold>0.85</v:threshold> …
    </param> …
</listen>
```

In this example, the <listen> object is extended for speaker verification that compares a user's voice against a cohort set. The events "onreco" and "onnoreco" are invoked when the voice passes or fails the test, respectively. As in the previous case, the extension must be decorated with URN that specifies the intended behavior of the document author. Being an extension, the document processor might not be able to discern the semantics implied by the URN natively. However, XML based protocols allow the processor to query and employ Web services that can either (1) transform the extended document segment into an XML schema the processor understands, or (2) perform the function described by the URN. By closely following XML standards, SALT documents fully enjoy the benefits of extensibility and portability of XML with SALT.

3.2 Language extensibility

In addition to component extensibility, the whole language of SALT can be enhanced with new functionality using XML. Communicating with other modalities, input devices, and advanced discourse and context management are just a few potential use for language-wise extension.

SALT message extension, or the <smex> element, is the standard conduit between a SALT document and the outside world. The message element takes <param> as its child element to forge a connection to an external component. Once a link is established, SALT document can communicate with the external component by exchanging text messages. The <smex> element has a "sent" attribute to which a text message can be assigned to. When its value is changed, the new value is regarded as a message intended for the external component and immediately dispatched. When an incoming message is received, the object places the message on a property called "received" and raises the "onreceive" event.

3.2.1 Telephony Interface

Telephones are one of the most important access devices for spoken language enabled Web applications. Call control functions play a central role in a telephony SALT application. The <smex> element in SALT is a perfect match with the telephony call standard known as ECMA 323. ECMA 323 defines the standard XML schemas for messages of telephony functions, ranging from simple call answering, disconnecting, transferring to switching functionality suitable for large call centers. ECMA 323 employs the same message exchange model as the design of <smex>. This allows SALT application to tap into a rich telephony call controls without needing a complicated SALT processor. As shown in (SALT 2002), sophisticated telephony applications can be authored in SALT in a very straightforward manner.

3.2.2 Advanced Context Management

In addition to devices such as telephones, SALT <smex> object may also be employed to connect to Web services or other software components to facilitate advanced discourse semantic analysis and context managements. Such capabilities, as

described in (Wang 2001), empower the user to realize the full potential of interacting with computers in natural language. For example, the user can simply say "Show driving directions to my next meeting" without having to explicitly and tediously instruct the computer to first look up the next meeting, obtain its location, copy the location to another program that maps out a driving route.

The customized semantic analysis can be achieved in SALT as follows:

```
<listen id="first_pass" …>
  <grammar src="…" /> <!-- basic grammar -->
  <bind targetElement="contextManager"
        targetAttribute="sent" value="/" />
  …
</listen>

<smex id="contextManager" …>
  <param xmlns:ws="WebServices">
    <ws:url>http://personal.com</ws:url>
    <ws:user>…</ws:user> …
  </param>
  <bind targetElement="realTarget" … />
</smex>
```

Here the <listen> element includes a rudimentary grammar to analyze the basic sentence structure of user utterance. Instead of resolving every reference (e.g. "my next meeting"), the result is cascaded to the <smex> element linked to a Web service specializing in resolving personal references.

Summary

In this paper, we describe the design philosophy of SALT in using the existing Web architecture for distributed multimodal dialog. SALT allows flexible and powerful dialog management by fully taking advantage of the well publicized benefits of XML, such as separation of data from presentation. In addition, XML extensibility allows new functions to be introduced as needed without sacrificing document portability. SALT uses this mechanism to accommodate diverse Web access devices and advanced dialog management.

References

Sadek M.D., Bretier P., Panaget F. (1997) *ARTIMIS: Natural dialog meets rational agency*. In Proc. IJCAI-97, Japan.

Allen J.F. (1995) Natural language understanding, 2nd Edition, Benjamin-Cummings, Redwood City, CA.

Cohen P.R, Morgan J., Pollack M.E (1989) Intentions in communications, MIT Press, Cambridge MA.

SALT Forum (2002) Speech Application Language Tags Specification, http://www.saltforum.org.

Wang K. (2000) *Implementation of a multimodal dialog system using extensible markup language,* In Proc. ICSLP-2000, Beijing China.

Aron B. (1991) *Hyperspeech: navigation in speech-only hypermedia*, in Proc. Hypertext-91, San Antonio TX.

Ly E., Schmandt C., Aron B. (1993) *Speech recognition architectures for multimedia environments*, in Proc. AVIOS-93, San Jose, CA.

Lau R., Flammia G., Pao C., Zue V. (1994) *Webgalaxy: Integrating spoken language and hyptertext navigation,* in Proc. EuroSpeech-97, Rhodes, Greece.

Sneff S., Hurley E., Lau R., Pao C., Schmid P., Zue V. (1998) *Galaxy-II: A reference architecture for conversational system development*, in Proc. ICSLP-98, Sydney Australia.

Rudnicky A., Xu W. (1999) An agenda-based dialog management architecture for spoken language systems, in Proc. ASRU-99, Keystone CO.

Lin B.-S, Wang H.-M, Lee L.-S (1999) *A distributed architecture for cooperative spoken dialog agents with coherent dialog state and history,* in Proc. ASRU-99, Keystone CO.

Bradshaw J.M. (1996) Software Agents, AAAI/MIT Press, Cambridge, MA.

Wang K (1998) An event driven dialog system, in Proc. ICSLP-98, Sydney Australia.

Wang K (2001) *Semantic modeling for dialog systems in a pattern recognition framework*, in Proc. ASRU 2001, Trento Italy.

Johnston M., Cohen P.R., McGee D., Oviatt S.L., Pittman J.A., Smith I (1997) *Unification based multimodal integration*, in Proc. 35th ACL, Madrid Spain.

Wang K (2001) *Natural language enabled Web applications*, in Proc. 1st NLP and XML Workshop, Tokyo, Japan.