VoiceXML and Voice Application Development

Deborah Dahl
Max Froumentin
February 2012

Victor de Boer (VUA)
March 2014

With examples from https://studio.tellme.com/vxml2/ovw/essentials.html

About Voice Applications

Applications are accessed with a telephone

The application provides information by voice

The user can either use the telephone keypad or speak to respond to the application

Motivation for Speech Applications

Users access Web sites from any telephone, anywhere, any time.

Speaking and listening are the natural usage modes for phones.

Easy to integrate with human telephone conversations.

Voice Application Languages

1. Proprietary software with GUI
2. Common languages (PHP, JS) with proprietary APIs. E.g., tropo
3. Proprietary scripts. E.g., Asterisk
4. THE standard: VoiceXML

Web and voice applications

Visual Web	Voice Web
The user access an application through a browser and URL	The user calls a telephone number
The browser makes an HTTP request to a server for an HTML page	The voice browser makes an HTTP request to a server for a VoiceXML page
The browser creates a visual page that the user interacts with by mouse and keyboard	The VoiceXML browser renders the VoiceXML page as a dialog. The user communicates with the page by speech or keypresses
The application occurs through space and time	The application proceeds through time

Limitations of Voice Applications

?

Limitations of Voice Applications

VUI conversational bandwidth is slower than GUI conversational bandwidth

All information from the application has to be spoken to the user

All information from the user to the application has to be spoken or communicated with a keypress

User input options are limited

Evaluating Application Ideas: Technology

Inbound or outbound calls

Input: for DTMF input, limited number of options at any point

Doesn’t need alphabetic input

Very noisy environments make it hard to hear system prompts

Doesn’t need graphical display

Limited amount of information to be presented if the user has to remember it

For speech recognition, things users can say have to be limited

Voice User Interface Design Concerns

The application takes place in time

The users can’t see what the application wants them to do

The users can’t see what they did

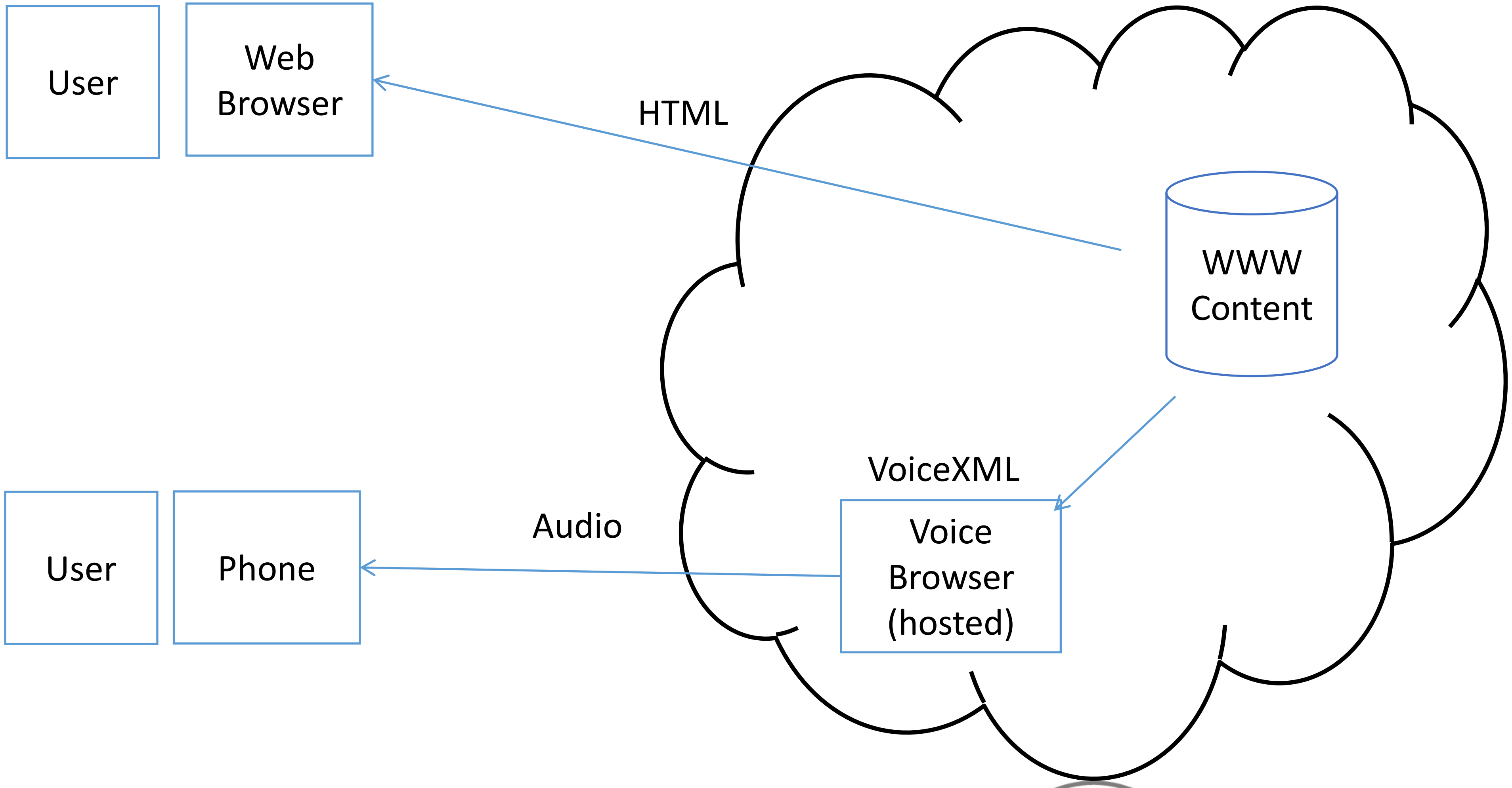
The users can’t see what the system thought they did

Simple VoiceXML Example on Voxeo Evolution

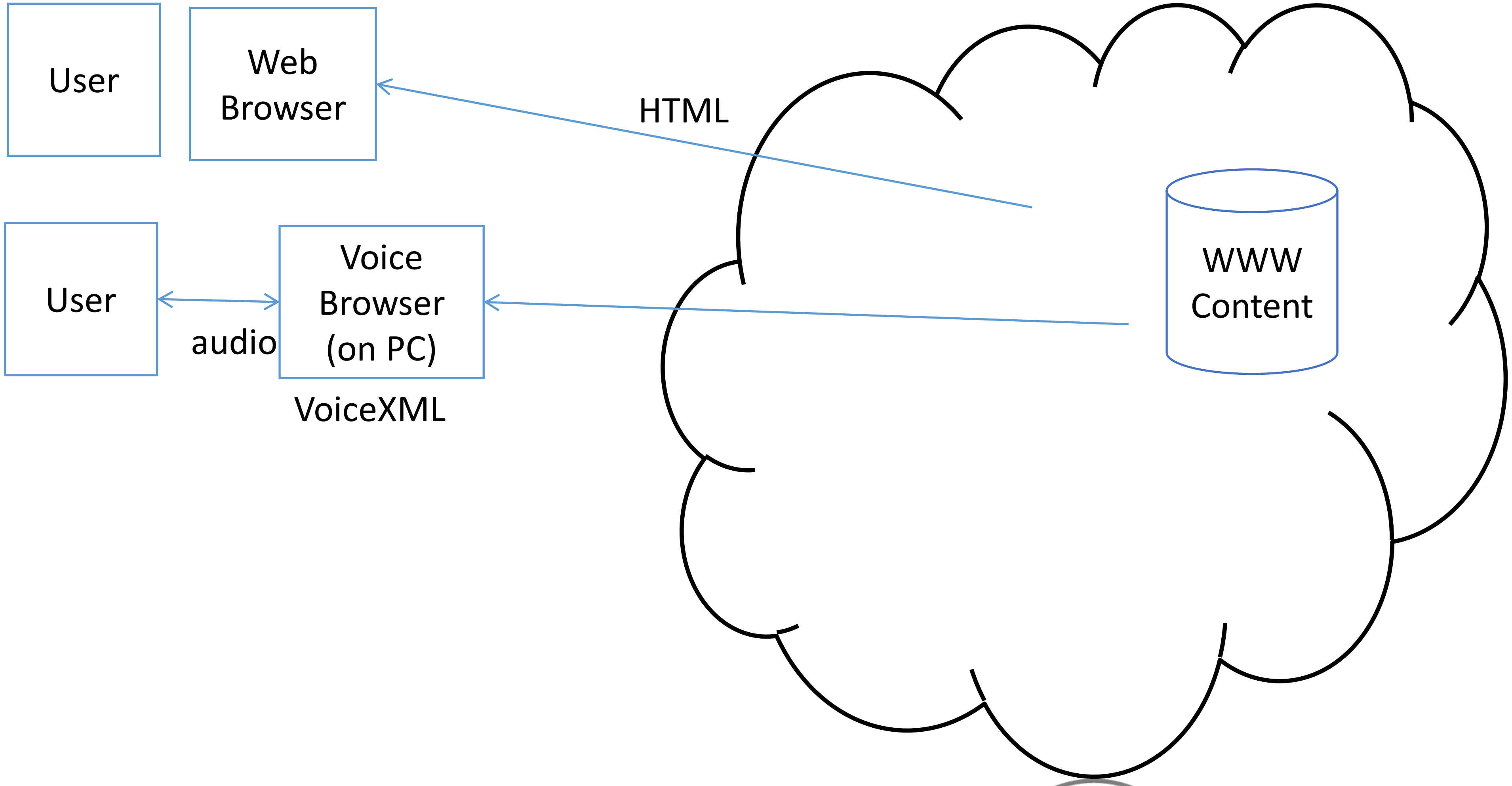
Test: +31 20 8082848   with pin 9990105221

```
<?xml version="1.0" encoding="UTF-8"?>
<vxml version = "2.1" >
  <form>
    <block>
      <prompt>
        Hello World!
      </prompt>
    </block>
  </form>
</vxml>
```

Voice on the Web



Voice on the Web



Basics of voice applications

Callflow– how the user progresses through the application from beginning to end

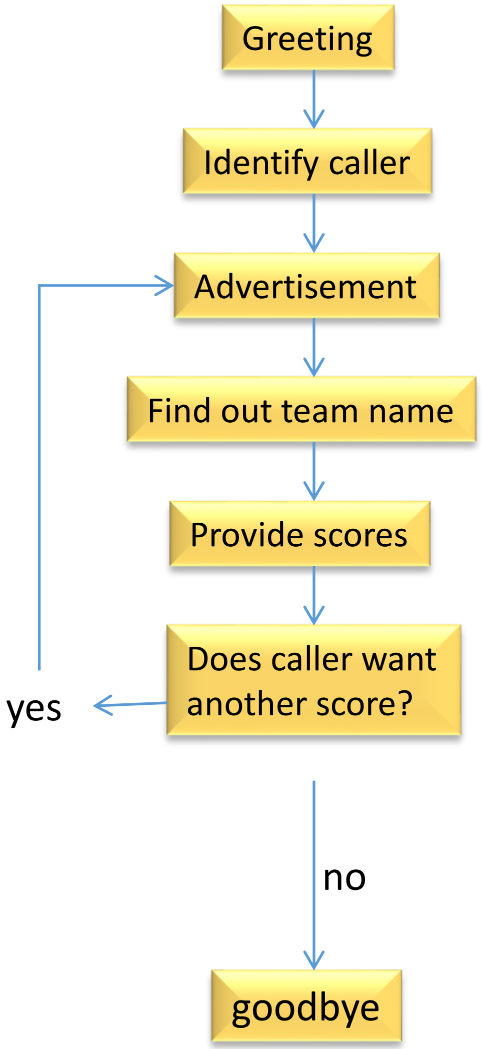
Dialogs – describe the sequence of what the system says and how the caller is expected to respond

System output: Prompts – audio files or text that the system speaks

User input: Grammars – describe how to interpret the caller’s speech or keypresses

High Level Callflow

Example, sports scores

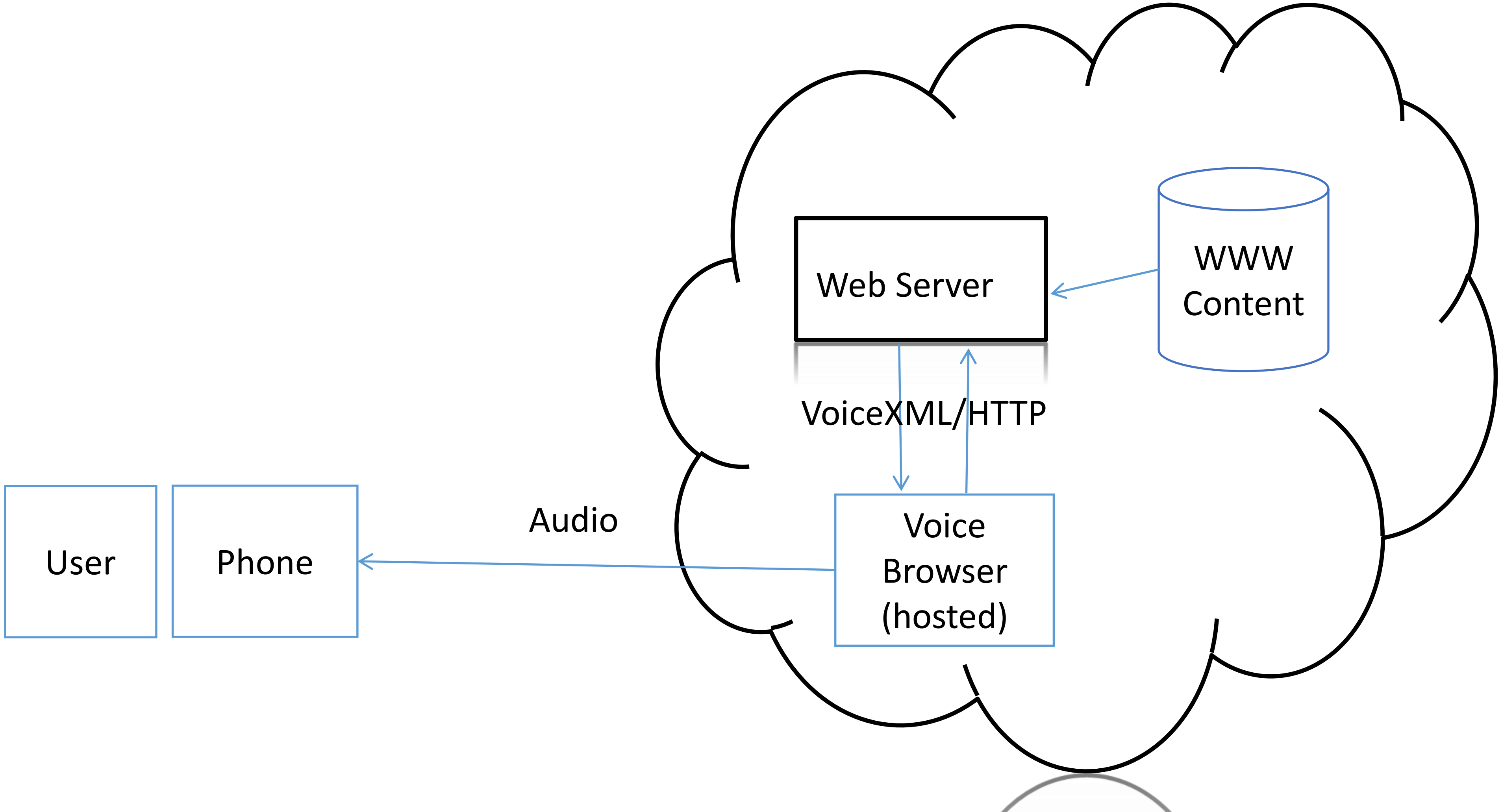


Low Level Callflow Design

State Name: Find out team name

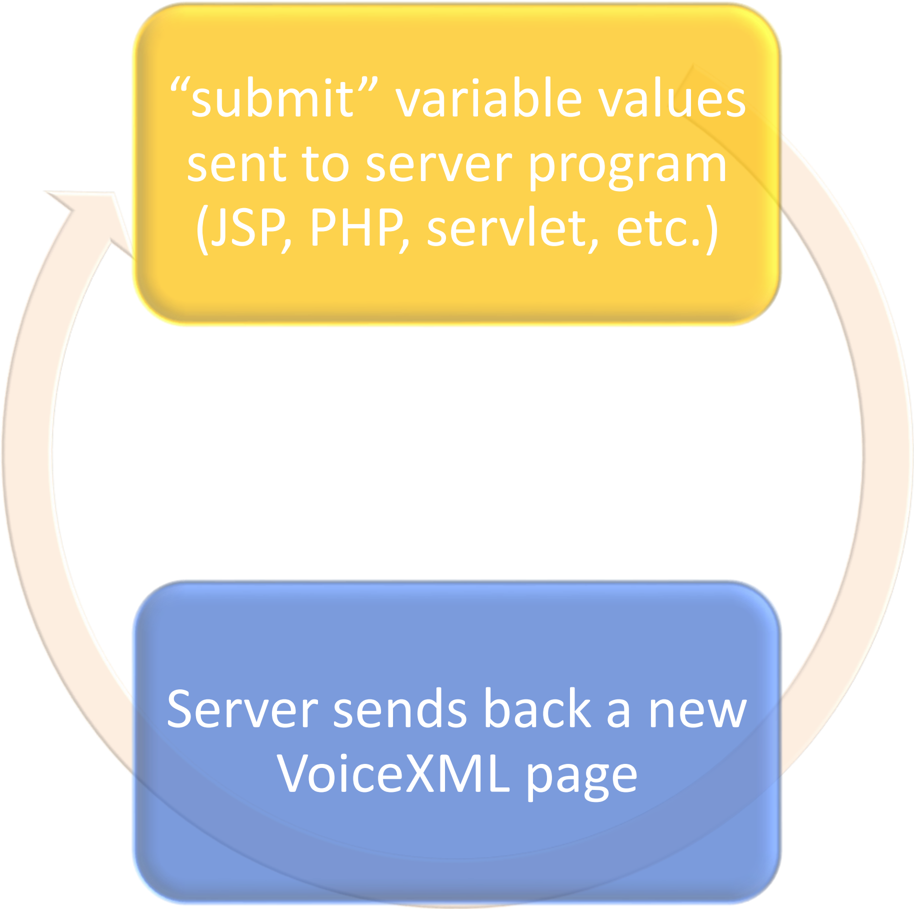
System	User	Actions	Next
For the Harambee Stars, press 1, for the AFC Leopards, press 2, for Mathare United, press 3, for all other teams press 4	1	Team= “Harambee Stars”	Goto “provide score”
	2	Team=“AFC Leopards”	
	3	Team=“Mathare United”	
	4		Goto “other teams”
	5	Throw error, “not a choice”	catch error, play Error message
Hidden choice	6	Transfer to operator	

Voice on the Web



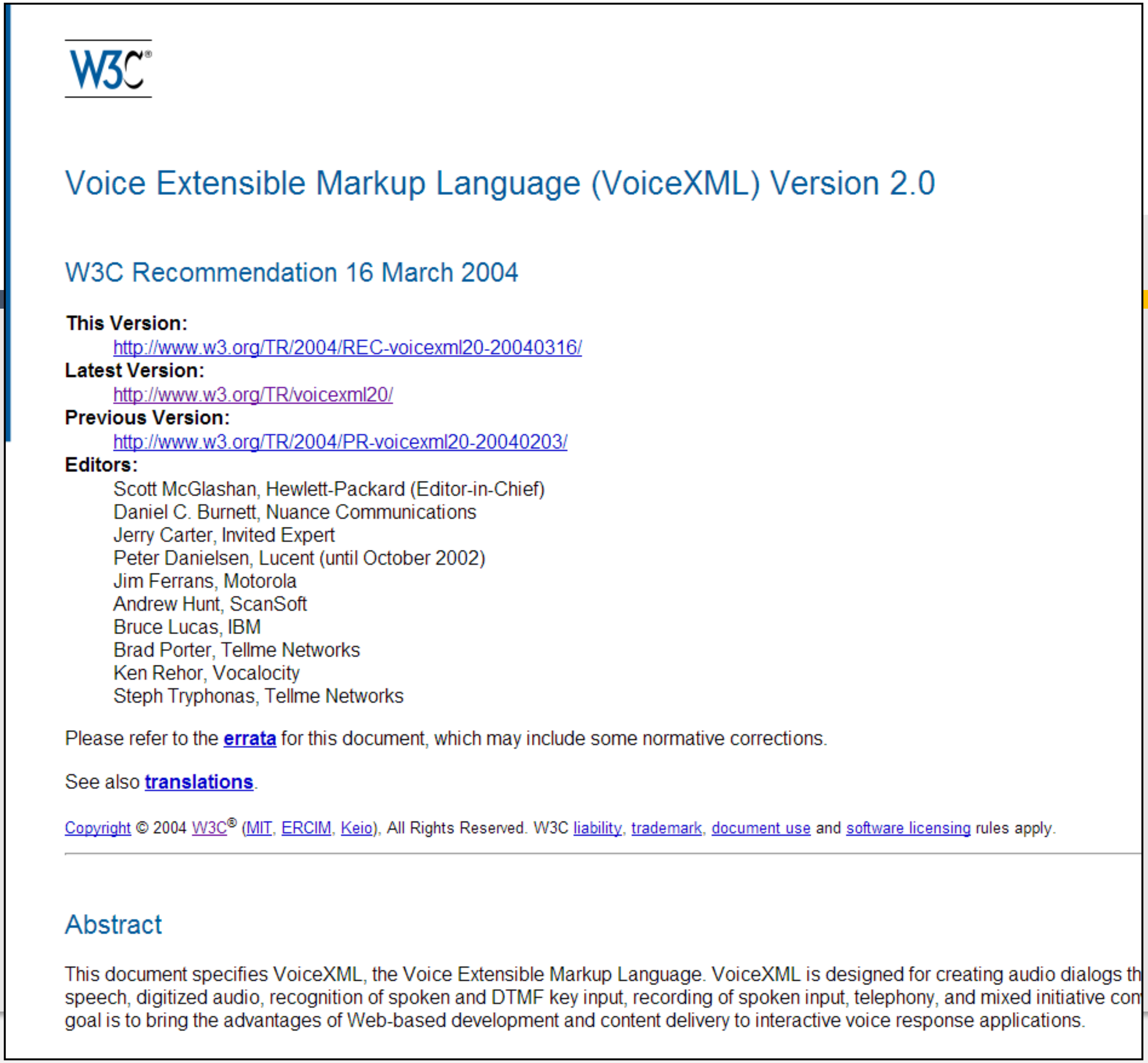
Server-side processing

The same as with the graphical web



“submit” variable values sent to server program (JSP, PHP, servlet, etc.)

Server sends back a new VoiceXML page



Voice Extensible Markup Language (VoiceXML) Version 2.0

W3C Recommendation 16 March 2004

This Version:
http://www.w3.org/TR/2004/REC-voicexml20-20040316/
Latest Version:
http://www.w3.org/TR/voicexml20/
Previous Version:
http://www.w3.org/TR/2004/PR-voicexml20-20040203/
Editors:
Scott McGlashan, Hewlett-Packard (Editor-in-Chief)
Daniel C. Burnett, Nuance Communications
Jerry Carter, Invited Expert
Peter Danielsen, Lucent (until October 2002)
Jim Ferrans, Motorola
Andrew Hunt, ScanSoft
Bruce Lucas, IBM
Brad Porter, Tellme Networks
Ken Rehor, Vocalocity
Steph Tryphonas, Tellme Networks

Please refer to the **errata** for this document, which may include some normative corrections.

See also **translations**.

Abstract

This document specifies VoiceXML, the Voice Extensible Markup Language. VoiceXML is designed for creating audio dialogs th speech, digitized audio, recognition of spoken and DTMF key input, recording of spoken input, telephony, and mixed initiative con goal is to bring the advantages of Web-based development and content delivery to interactive voice response applications.

VoiceXML is an XML language

XML = eXtensible Markup Language

Elements are surrounded by tags

```
<prompt>Welcome to the voice system </prompt>
```

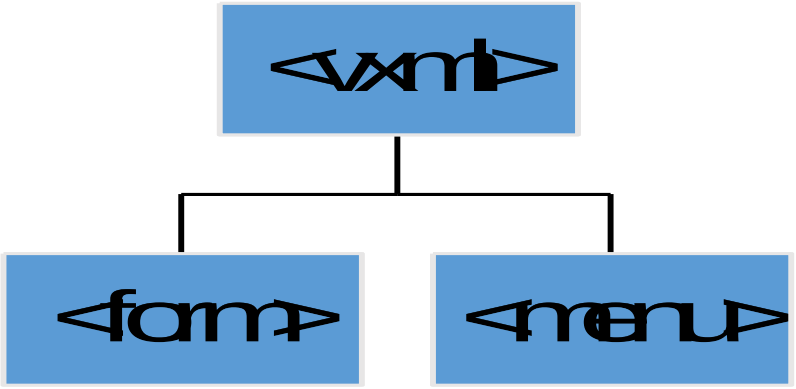
Elements may be nested

```
<prompt>
    Welcome to Ajax Travel <break/>
    we have the cheapest fares
</prompt>
```

Elements may have attributes

```
<choice next="#boat">
<grammar type="application/grammar+xml" version="1.0"
    root = "by_boat" src = “boat.grxml”>
```

VoiceXML Top-level constructs



<vxml>

<form>

<menu>

A **menu** presents the user with a choice of options and the transitions to another dialog state based upon the users selection.

A **form** defines an interaction that collects values for each of the ***fields*** in the form. Each field may specify a prompt, the expected input, and evaluation rules. (cf. HTML form)

Parts of a dialog

- The system says something using the <prompt> element
- The user replies with speech or keypresses
- This sequence of steps is called a *turn*
- The system does something based on the user’s input
  - submits data to a server
  - transitions to another dialog
  - transfers a call
  - executes a script

Example: Hello World (again)

```
<?xml version="1.0" encoding="UTF-8"?>

<vxml version = "2.1" >
    <form>
      <block>
        <prompt>
          Hello World!

        </prompt>
      </block>
    </form>

</vxml>
```

Example: Hello World (again)

```
<?xml version="1.0" encoding="UTF-8"?>

<vxml version = "2.1" >
     <form>
       <block>
         <prompt>

                      <audio src=“hello.wav”/>
         </prompt>
       </block>
     </form>

</vxml>
```

System output

- Recorded Audio: sounds natural, can be familiar voice or language.
- Concatenation (eg, Radio Marché)
- Speech Synthesis: good when lots of possible different messages. Still sounds funny (but improving). ~20 languages

Example

```
<?xml version="1.0" encoding="UTF-8"?>
<vxml version = "2.1" >
  <form>
    <block>
      <prompt>
        Hello, I’m spoken by a TTS
        <audio src=“http://…/human.wav”/>
      </prompt>
    </block>
  </form>
</vxml>
```

Goto

```
<vxml version="2.1" xmlns="http://www.w3.org/2001/vxml">
<form>
  <block>
        Goodbye, world!
        <goto next="document2.vxml"/>
 </block>
</form>
```

```
Or <goto next="#infinity"/>
```

User input: Keypresses

- Telephones with a physical or soft keypad generate tones for each key that communicate with an application
- The tones are called DTMF tones (Dual Tone Multiple Frequency)

Useful for: noisy environments, privacy, entering numbers, languages for which there isn’t a speech recognizer

DTMF 1 2 3 4 5 6 7 8 9 0 * #

User input: Speech recognition

The user communicates with the application by speaking

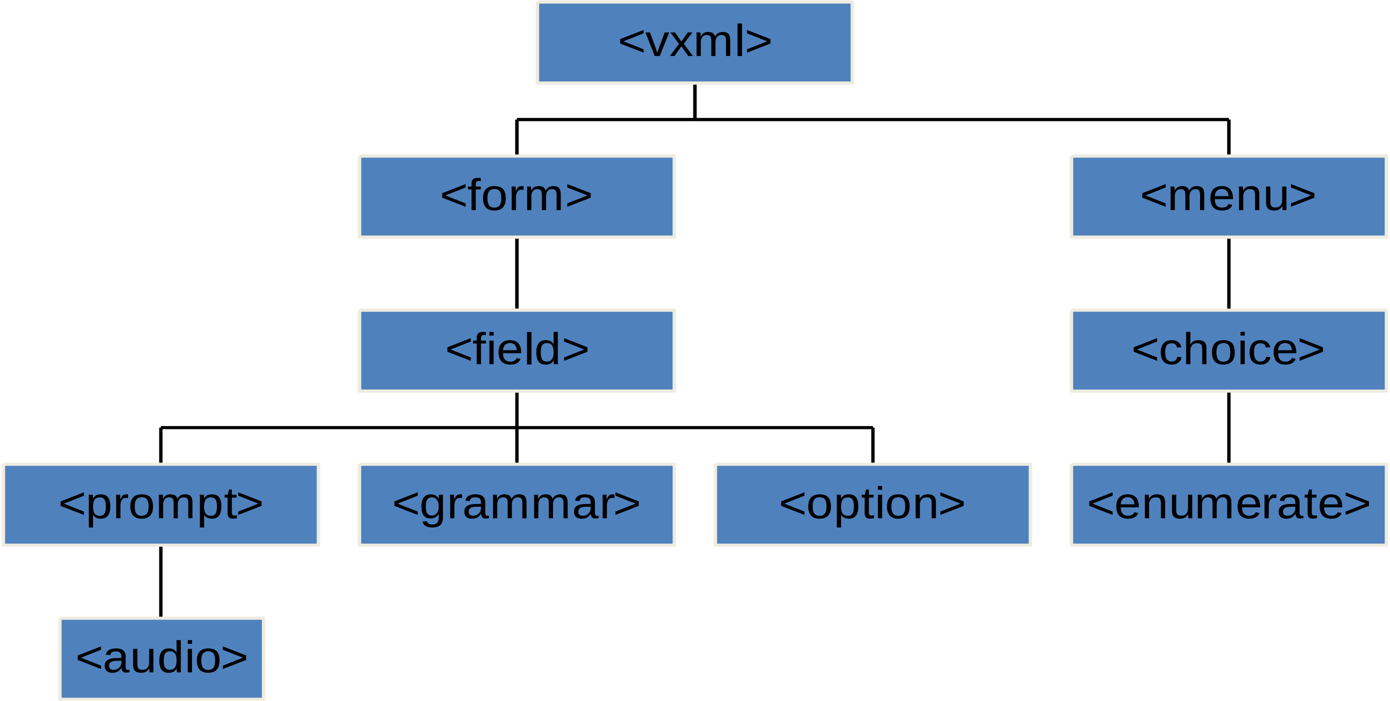
An automatic speech recognition (ASR) system analyzes the user’s speech and determines the words that were spoken

Less accurate than DTMF tone recognition

Speech recognition compared to DTMF

- DTMF is useful for:
  - Noisy environments, because it’s more accurate than speech recognition
  - Privacy, because people can’t interpret it
  - Entering numbers
  - Languages for which there isn’t a speech recognizer because it’s language-independent
  - We will focus on DTMF in this class
- Speech recognition is useful for:
  - Meaningful lists longer than 10 items
  - Obtaining multiple pieces of information at once
  - Reducing the memory load on the users

VoiceXML Language High Level Elements



Field

```
<vxml version="2.1"
 xmlns="http://www.w3.org/2001/vxml">
  <form>
   <field name="famous" type="boolean">
    <prompt>
    would you like to be famous?
    </prompt>

    <filled>
    got it!
    <if cond="famous">
     let's schedule an audition
     <goto next="schedule.vxml" />
    <else />
     infamous is a reasonable
     alternative.
     <goto next="infamous.vxml" />
    </if>
    </filled>
   </field>
  </form>
</vxml>
```

Built in grammar (“yes”, “ok”, “no”, “nope”)

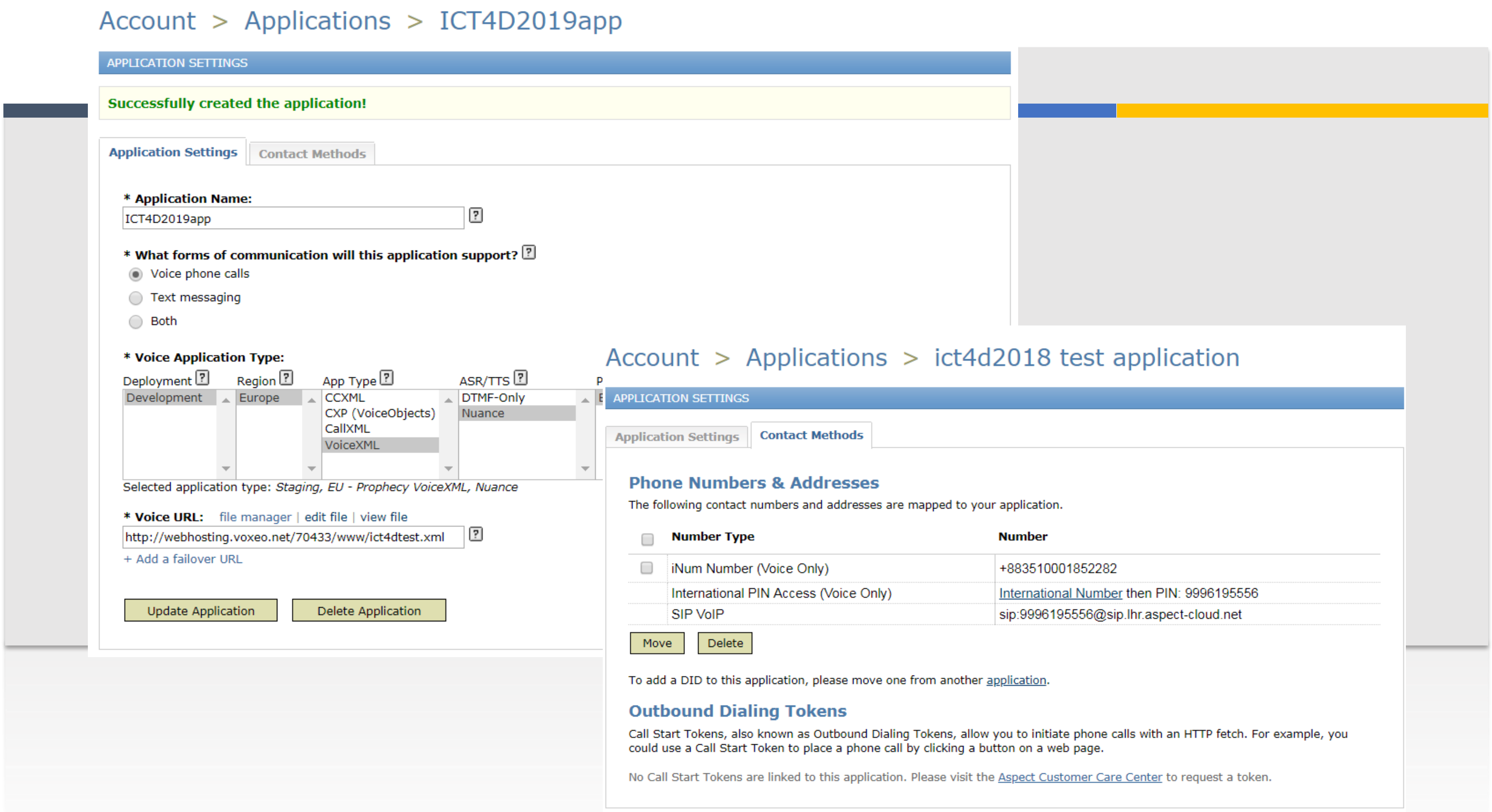
```
<noinput>
   Sorry I didn't hear you.
   For famous press 1.
   For infamous press 2.
 </noinput>

 <nomatch>
   Sorry I didn't get that.
   To be famous say yes.
   Otherwise, say no.
 </nomatch>
```

Intermission

Voxeo Evolution

Evolution.voxeo.com



Account > Applications > ICT4D2019app

APPLICATION SETTINGS

Successfully created the application!

*** Application Name:**

ICT4D2019app

*** What forms of communication will this application support?**

● Voice phone calls

○ Text messaging

○ Both

*** Voice Application Type:**

Deployment	Region	App Type	ASR/TTS
Development	Europe	CCXML CXP (VoiceObjects) CallXML VoiceXML	DTMF-Only Nuance

Selected application type: *Staging, EU - Prophecy VoiceXML, Nuance*

*** Voice URL:** file manager | edit file | view file

http://webhosting.voxeo.net/70433/www/ict4dtest.xml

+ Add a failover URL

Account > Applications > ict4d2018 test application

APPLICATION SETTINGS

Phone Numbers & Addresses

The following contact numbers and addresses are mapped to your application.

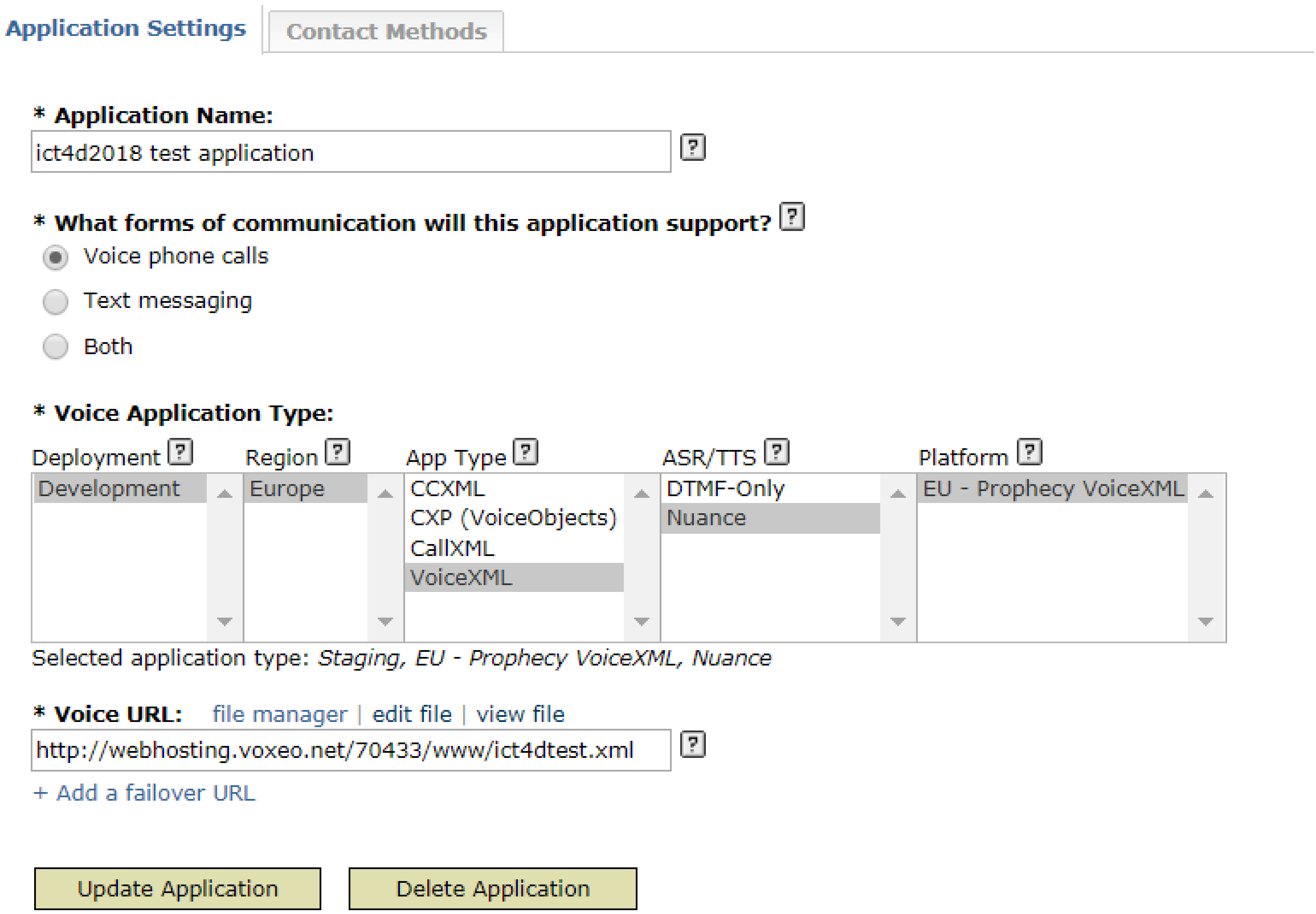
□	Number Type	Number
□	iNum Number (Voice Only)	+883510001852282
	International PIN Access (Voice Only)	International Number then PIN: 9996195556
	SIP VoIP	sip:9996195556@sip.lhr.aspect-cloud.net

To add a DID to this application, please move one from another application.

Outbound Dialing Tokens

Call Start Tokens, also known as Outbound Dialing Tokens, allow you to initiate phone calls with an HTTP fetch. For example, you could use a Call Start Token to place a phone call by clicking a button on a web page.

No Call Start Tokens are linked to this application. Please visit the Aspect Customer Care Center to request a token.



* **Application Name:** ict4d2018 test application

* **What forms of communication will this application support?**

● Voice phone calls

○ Text messaging

○ Both

* **Voice Application Type:**

Deployment	Region	App Type	ASR/TTS	Platform
Development	Europe	CCXML	DTMF-Only	EU - Prophecy VoiceXML
		CXP (VoiceObjects)	Nuance	
		CallXML		
		VoiceXML		

Selected application type: *Staging, EU - Prophecy VoiceXML, Nuance*

* **Voice URL:** http://webhosting.voxeo.net/70433/www/ict4dtest.xml

PHP!

```
<vxml version=“2.1”>
  <form>
    <block>
<prompt>Hello <?= $_POST[‘name’]?></prompt>

…
```

(You can use Ruby/Python/node.js/ASP, or any server-side language)