Средства распознавания и синтеза речи

На примере Microsoft Speech Platform

Microsoft Speech Platform 11

- Набор средств для разработки приложений с поддержкой синтеза и распознавания речи – чат-ботов, онлайн-консультантов, приложений с голосовым интерфейсом.
- Поддерживает managed-code и native-code application programming interfaces (APIs).
- Не является интерфейсом к сетевым сервисам, может использоваться без доступа к Интернету.
- Требует установки Speech Platform Runtime 11, также имеет набор средств разработки – SDK.

Microsoft Speech Platform 11

- Собственно сама платформа является «устаревшей», и, по всей видимости, не поддерживается.
- При этом позволяет использовать стандарты W3C для реализации грамматик, лексиконов и прочего.
- Для работы требуется установка Runtime и языков – отдельно для синтеза (TTS) и для распознавания (SR).
- Разработка поддерживается на Vista и далее, развёртывание – на Windows Server 2003 и выше.

Microsoft Speech Platform 11

- Собственно сама платформа является «устаревшей», и, по всей видимости, не поддерживается.
- При этом позволяет использовать стандарты W3C для реализации грамматик, лексиконов и прочего.
- Для работы требуется установка Runtime и языков – отдельно для синтеза (TTS) и для распознавания (SR).
- Разработка поддерживается на Vista и далее, развёртывание – на Windows Server 2003 и выше.

Синтез речи – этапы

- Синтезатор должен выполнить анализ текста – выделить слова, части речи, собственные имена. Определить грамматические атрибуты – время, число, падеж и прочее.
- Предварительный анализ – грамматический морфология, синтаксис, семантика.
- Второй этап – синтез звука. Используются базы данных, хранящие различные комбинации звуков – чем они больше, тем лучше синтез.
- Интонации, к сожалению, пока даются с трудом.

Синтез речи – этапы

- Синтезатор должен выполнить анализ текста – выделить слова, части речи, собственные имена. Определить грамматические атрибуты – время, число, падеж и прочее.
- Предварительный анализ – грамматический морфология, синтаксис, семантика.
- Второй этап – синтез звука. Используются базы данных, хранящие различные комбинации звуков – чем они больше, тем лучше синтез.
- Интонации, к сожалению, пока даются с трудом.

Простейший синтез

- Проговаривает фразу, без учёта особенностей, с использованием установленного голоса.

```
using System;
using System.Collections.Generic;
using System.Linq;
using System.Text;
using Microsoft.Speech.Synthesis;

namespace SimpleSpeak
{
  class Program
  {
    static void Main(string[] args)
    {
      SpeechSynthesizer s = new SpeechSynthesizer();
      s.Speak("Hello. My name is Microsoft Server Speech Text to Speech Voice (en-US, Helen).");
    }
  }
}
```

Голоса

- Хранят информацию о языке, поле, возрасте и проч.

Name	Description
AdditionalInfo	Gets additional information about the voice.
Age	Gets the age of the voice.
Culture	Gets the culture of the voice.
Description	Gets the description of the voice.
Gender	Gets the gender of the voice.
Id	Gets the ID of the voice.
Name	Gets the name of the voice.
SupportedAudioFormats	Gets the collection of audio formats that the voice supports.

PromptBuilder

- Позволяет конструировать сложные фразы и предложения, с указанием отдельных элементов.
- По умолчанию создаётся для текущих установок CultureInfo, может быть изменён.
- Может включать отдельные фрагменты аудио, а также SSML.
- Для текста можно указывать подсказки – как сказать, а также темп, акцент и громкость.
- Закладки – для контроля времени.

PromptBuilder

- Позволяет конструировать сложные фразы и предложения, с указанием отдельных элементов.
- По умолчанию создаётся для текущих установок CultureInfo, может быть изменён.
- Может включать отдельные фрагменты аудио, а также SSML.
- Для текста можно указывать подсказки – как сказать, а также темп, акцент и громкость.
- Закладки – для контроля времени.
- Текст можно разделять на абзацы и предложения.

Использование SSML

- Язык разметки, соответствует W3C **Speech Synthesis Markup Language (SSML) Version 1.0**
- Текущая версия – 1.1, сентябрь 2010.
- Примерно тогда же рекомендован стандарт **Voice Extensible Markup Language (VoiceXML) 3.0**
- Позволяет размечать и описывать структуру речи примерно так же, как с содержимым веб-страниц.

Стандарты W3C в области речи

Completed Work

2015-09-01	State Chart XML (SCXML): State Machine Notation for Control Abstraction	Recommendation
2011-07-05	Voice Browser Call Control: CCXML Version 1.0	Recommendation
2010-09-07	Speech Synthesis Markup Language (SSML) Version 1.1	Recommendation
2008-10-14	Pronunciation Lexicon Specification (PLS) Version 1.0	Recommendation
2007-06-19	Voice Extensible Markup Language (VoiceXML) 2.1	Recommendation
2007-04-05	Semantic Interpretation for Speech Recognition (SISR) Version 1.0	Recommendation
2004-09-07	Speech Synthesis Markup Language (SSML) Version 1.0	Recommendation
2004-03-16	Speech Recognition Grammar Specification Version 1.0	Recommendation
2004-03-16	Voice Extensible Markup Language (VoiceXML) Version 2.0	Recommendation
2015-08-11	DOM Event I/O Processor for SCXML	Group Note
2015-08-11	XPath Data Model for SCXML	Group Note
2009-12-08	Mobile Web for Social Development Roadmap	Group Note
2005-05-26	SSML 1.0 say-as attribute values	Group Note
1998-01-28	Voice Browsers	Group Note

Drafts

2012-03-20	CSS Speech Module	Candidate Recommendation
2010-12-16	Voice Extensible Markup Language (VoiceXML) 3.0	Working Draft
2008-08-08	Voice Extensible Markup Language (VoiceXML) 3.0 Requirements	Working Draft
2007-06-11	Speech Synthesis Markup Language Version 1.1 Requirements	Working Draft

Пример

```
<?xml version="1.0"?>
<speak version="1.1" xmlns="http://www.w3.org/2001/10/synthesis"
      xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
      xsi:schemaLocation="http://www.w3.org/2001/10/synthesis
            http://www.w3.org/TR/speech-synthesis11/synthesis.xsd"
      xml:lang="en-US">
  The French word for cat is <w xml:lang="fr">chat</w>.
  He prefers to eat pasta that is <lang xml:lang="it">al dente</lang>.
</speak>
```

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<speak version="1.0"
 xmlns="http://www.w3.org/2001/10/synthesis"
 xml:lang="en-US">
  <s>
    His name is Mike <phoneme alphabet="x-microsoft-ups" ph="JH AU"> Zhou
</phoneme>
  </s>
</speak>
```

Этапы обработки SSML

1. Парсинг XML.
2. Структурный анализ текста – абзацы и предложения. Возможна поддержка тегами <s> и <p>.
3. Нормализация текста. «$200» – «two hundred dollars». «1/2» – «half», «January second», «February first», «one of two»? Результат – токены, обычно слова. Поддержка тегами <say-as> и <sub>. 今日は может произноситься как きょうは («kyou wa» = "today") или こんにちは («konnichiwa» = "hello") при поддержке Kanji и kana.
4. Text-to-phoneme conversion. Токены в последовательность фонем. Английский – 45, Гавайский – 12-18, некоторые языки до 100. Проблемы – «I will read the book» or «I have read the book». Нестандартные произношения заимствований и иностранных слов: «Caius College» (произносится «keys college») and President Tito (произносится «sutto»), the president of the Republic of Kiribati (произносится «kiribass»). Поддержка тегами <phoneme>, <lexicon>, <sub> и прочими.

Этапы обработки SSML

5. Анализ просодии – системы единиц звучания (ударение, тон, интонация). Темп, паузы, ударения и прочее. Поддержка тегами <emphasis>, <break>, <prosody>.
6. Генерация звука. Сложный этап, управлять можно голосами (для Microsoft Speech Platform по одному голосу на язык примерно). Также можно вставлять фрагменты аудио. Поддержка тегами <voice> и <audio>.

<s> His name is Mike <phoneme alphabet="x-microsoft-ups" ph="JH AU"> Zhou </phoneme> </s>

<s> Your order for <prosody pitch="+1st" rate="-10%" volume="90"> 8 books </prosody> will be shipped tomorrow. </s>

Лексиконы и фонетические алфавиты

- Лексиконы – словари произношений слов и коротких фраз. Обычно есть один лексикон по умолчанию, может быть модифицирован.
- Фонетические алфавиты позволяют явно указать звучание отдельных элементов.
- Поддерживаются три фонетических алфавита – International Phonetic Alphabet (IPA), Universal Phone Set (UPS от Microsoft) и Speech API (SAPI) Phone Set – специализированный для некоторых языков.

Лексиконы и фонетические алфавиты

- Лексиконы – словари произношений слов и коротких фраз. Обычно есть один лексикон по умолчанию, может быть модифицирован.
- Фонетические алфавиты позволяют явно указать звучание отдельных элементов.
- Поддерживаются три фонетических алфавита – International Phonetic Alphabet (IPA), Universal Phone Set (UPS от Microsoft) и Speech API (SAPI) Phone Set – специализированный для некоторых языков.

Фонетические алфавиты

A	Álpha
B	Brávo
C	Chárlie
D	Délta
E	Écho
F	Fóxtrot
G	Gólf
H	Hotél
I	Índia
J	Júliet
K	Kílo
L	Líma
M	Mike
N	Novémber
O	Óscar
P	Pápa
Q	Quebéc
R	Rómeo
S	Siérra
T	Tángo
U	Úniform
V	Víctor
W	Whísky
X	X-ray
Y	Yánkee
Z	Zúlu

3	Tree
4	Fower
5	Fife
9	Niner

А	Анто́н (Анто́н)
Б	Бори́с
В	Васи́лий
Г	Григо́рий (Гали́на)
Д	Дми́трий
Е	Еле́на
Ё	Еле́на (Ёлка)
Ж	Же́ня (Жук)
З	Зинаи́да (Зо́я)
И	Ива́н
Й	Ива́н кра́ткий (Йот)
К	Константи́н (Килова́тт)
Л	Леони́д
М	Михаи́л (Мари́я)
Н	Никола́й
О	Ольга
П	Па́вел
Р	Рома́н (Ра́дио)
С	Семён (Серге́й)
Т	Татья́на (Тама́ра)
У	Улья́на
Ф	Фёдор
Х	Харито́н
Ц	Ца́пля (Центр)
Ч	Челове́к
Ш	Шу́ра
Щ	Щу́ка
Ъ	Твёрдый знак
Ы	Еры (и́грек)
Ь	Мя́гкий знак (Знак)
Э	Эхо (Эмма)
Ю	Юрий
Я	Яков

Фонетические алфавиты

UPS	IPA	Unicode	SAPI ID	IPA Description	IpaASCII	X-SAMPA	Example	Language
I	i	U+0069	0069	{hgh,fnt,unr,vwl}	I	i	feel	English
Y	y	U+0079	0079	{hgh,fnt,rnd,vwl}	Y	y	du	French
IX	ɨ	U+0268	0268	{hgh,cnt,unr,vwl}	i"	1		
YX	ʉ	U+0289	0289	{hgh,cnt,rnd,vwl}	U"	}		
UU	ɯ	U+026F	026F	{hgh,bck,unr,vwl}	u-	M		
U	u	U+0075	0075	{hgh,bck,rnd,vwl}	u	u	too	English
IH	ɪ	U+026A	026A	{smh,fnt,unr,vwl}	I	I	fill	English
YH	ʏ	U+028F	028F	{smh,fnt,rnd,vwl}	I.	Y	hübsch	German