IBM Text-to-Speech SSML Programming Guide

December 2008

A form for readers' comments appears at the back of this publication. If the form has been removed, address your comments to:

International Business Machines Corporation
Department MMOA
P.O. Box 12195
Research Triangle Park, North Carolina
27709-2195

When you send information to IBM®, you grant IBM a nonexclusive right to use or distribute the information in any way it believes appropriate without incurring any obligation to you.

Contents

IBM Text-to-Speech SSML Programming Guide

This guide provides information about using Speech Synthesis Markup Language (SSML) to incorporate IBM Text-to-Speech technology into applications.

About using IBM Text-to-Speech and SSML

This guide describes the programming interfaces available for developers to take advantage of these features within their applications.

This guide is also available in Portable Document Format (PDF). To read it, you must have the Adobe Acrobat Reader installed on your computer. Using IBM Text-to-Speech Technology and Speech Synthesis Markup Language

Who should use this guide

This guide can help if you are a software developer interested in writing applications that use IBM Text-to-Speech technology. This document describes the use of IBM Text-to-Speech technology for beginning to advanced software engineers.

Guide topics

This guide presents the following main topics.

- SSML

  This topic describes the use of SSML to control speech synthesis and text processing parameters.
- Symbolic phonetic representations (SPRs)

  This topic describes the use of special phonetic symbols to customize pronunciations in IBM Text-to-Speech.

Typographical conventions

The following typographical conventions are used throughout this document to facilitate reading and comprehension. They are outlined in the following list.

`Monospace font`
: Applies to code samples, including XML, and to file and directory names.

Bold
: Applies to function and callback names, and to data types, including structures and enumerations.

Italics
: Applies to parameter and structure member names, sample text, and the introduction of new terms.

UPPERCASE
: Applies to property, enumerator, mode, and state names.

Introduction to SSML

The VoiceXML 2.0 specification adopted Speech Synthesis Markup Language (SSML) as the standard markup language for speech synthesis.

SSML provides developers of speech applications a standard way in which to control speech synthesis and text processing parameters. SSML enables developers to specify pronunciations, volume, pitch, speed, and so on.

The SSML standard

This section describes SSML support in IBM Text-to-Speech System. The implementation is based on the Speech Synthesis Markup Language Version 1.0, recommended by W3C on September 7, 2004, which can be found at http://www.w3.org/TR/speech-synthesis/

The IBM Text-to-Speech System implements this specification, with the following exception:

The following `prosody` elements are not supported: `Contour` and `Duration`. The `rate` element is supported but the implementation varies from the current W3C specification.

SSML parsing

The SSML processor silently ignores unsupported tags. The text contained inside an unsupported **<say-as>** tag is synthesized as-is; that is, only the tag is ignored.

If the syntax of the input text is not legal, the SSML processor returns and logs an error.

SSML language support

The level of SSML support is language dependent. The topic ″SSML language support″ provides detailed information about what is supported for each language.

Related information

SSML language support
Support for some SSML tags is language-specific. This topic summarizes the supported languages and the SSML features supported for each language.

SSML phonemes

SSML supports two phonetic alphabets: the International Phonetic Alphabet (IPA) and the IBM TTS phonetic alphabet.

The optional alphabet attribute of the phoneme element specifies the phonetic alphabet. If the alphabet is not specified, the default IBM TTS phonetic alphabet is used.

International Phonetic Alphabet example

The following example shows the US English phonetic pronunciation of ″tomato″ using the IPA:

```
<phoneme alphabet="ipa" ph="t&#x259;&#x2C8;me&#x26A;.&#x27E;&#x28A;"> tomato </phoneme>
```

The following example shows the pronunciation of ″tomato″ using the character codes for the IPA symbols. If you have encoding to support the IPA symbols, you can these symbols instead:

```
<phoneme alphabet="ipa" ph="tˈmeɪ.ɾoʊ"> tomato </phoneme>
```

For more information about the IPA, refer to: http://www2.arts.gla.ac.uk/IPA/index.html.

IBM TTS phonetic alphabet example

The following example shows the US English phonetic pronunciation of ″tomato″ using the IBM TTS phonetic alphabet:

```
<phoneme alphabet="ibm" ph=".0tx.1me.0Fo"> tomato </phoneme>
```

The pronunciation of ″tomato″ is given in a notation called Symbolic Phonetic Representation (SPR). The SPR for a word is the phonetic spelling used by IBM TTS to represent the pronunciation of a word. An SPR represents the sounds of a word, how these sounds are divided into syllables, and which syllables receive stress.

For more information about SPRs, including tables of SPRs for several languages, see “Introduction to symbolic phonetic representations” on page 11 in this guide.

SSML language support

Support for some SSML tags is language-specific. This topic summarizes the supported languages and the SSML features supported for each language.

Language-specific code is required to implement some of the SSML tags, such as `<say-as>` for dates. This version of the Text-to-Speech engine supports the language-dependent SSML tags for a subset of the languages.

Table of supported languages and SSML features

The following table summarizes the features that are supported for each language.

Feature	**En_US En_UK Gr_GR**	**Ja_JP**	**ZH_CN**	**La_SP**	**Fr_CA**
Speak	+	+	+	+	+
Paragraph	+	+	+	+	+
Sentence	+	+	+	+	+
Say-as					
digits	+	+	+	+	+
letters	+	+	+	+	+
date	+	+	+		
Number	+	+	+		
ordinal	+	+	+		
cardinal	+	+	+		
telephone	+	+	+		
telephone w/ punctuation	+	+	+		
vxml:boolean	+	+	+		
vxml:date	+	+	+		
vxml:digit	+	+	+	+	+
vxml:currency	+	+	+		

Feature	En_US En_UK Gr_GR	Ja_JP	ZH_CN	La_SP	Fr_CA
vxml:number	+	+	+		
vxml:phone	+	+	+		
vxml:time	+	+	+		
Phoneme					
ibm	+	+	+	+	+
ipa	+	+			
Sub	+	+	+	+	+
Voice	+	+		+	+
name	+	+	+	+	+
Break	+	+	+	+	+
Prosody					
pitch	+				+
range	+				+
rate	+				+
volume	+	+	+	+	+
Audio	+	+	+	+	+
Mark	+	+	+	+	+
Lexicon	+			+	+

SSML tags

This section gives a list of SSML tags and examples of how those tags are used.

<speak>

This is the root element for SSML documents. Valid attributes are:

xml:lang

This is a required attribute specifying the language. Accepted values are at http://www.ietf.org/rfc/rfc3066.txt

xml:base

This is an optional attribute specifying the base URI to use for resolving relative paths.

version

This is a required attribute specifying the SSML Specification. The accepted value is ″1.0″.

Example:

```
<speak xml:lang="En-US" version="1.0" xml:base="http://
www.myfileserver.com/mydir">text to be spoken</speak>
```

<paragraph> or <p> or <sentence> or <s>

These are optional tags that can be used to give text structure hints to the TTS system. The only valid attribute is **xml:lang**, which does allow values to be placed even though language switching is not supported.

Example:

```
<speak xml:lang="En-US" version="1.0"
<paragraph>
<sentence>Text within a sentence tag.</sentence>
<s>More text.</s>
</paragraph>
</speak>
```

Note: If the enclosed text in an SSML <sentence> or <paragraph> tag does not end with an end-of-sentence punctuation character (like a period), a longer than normal pause is added to the synthesized audio for this text.

<say-as>

The say-as tag allows the author to indicate information on the type of text contained within the tag and to help specify the level of detail for rendering the text. The required attribute for this tag is **interpret-as** . There are two optional attributes, **format** and **detail**, which are only used with particular values within the **interpret-as** attribute. These optional attributes are illustrated within the entries for their associated values.

letters This value spells out the characters in a given word within the enclosed tag.

Example (This will spell out ″HELLO″):

```
<speak xml:lang="En-US" version="1.0">
<say-as interpret-as="letters">Hello</say-as>
</speak>
```

digits This value spells out the digits in a given number within the enclosed tag.

Example (This will spell out ″123456″):

```
<speak xml:lang="En-US" version="1.0">
<say-as interpret-as="digits">123456</say-as>
</speak>
```

vxml:digits

This value performs the same function as the **digits** value.

```
<speak xml:lang="En-US" version="1.0">
<say-as interpret-as="vxml:digits">123456</say-as>
</speak>
```

date This value will speak the date within the enclosed tag, using the format given in the associated **format** attribute. The **format** attribute is required for use with the date value of interpret-as, but if **format** is not present, the engine will still attempt to pronounce the date.

Example (This gives a list of dates in all the various formats:)

```
<speak xml:lang="En-US" version="1.0">
<say-as interpret-as="date" format="mdy">12/17/2005</say-as>
<say-as interpret-as="date" format="ymd">2005/12/17</say-as>
<say-as interpret-as="date" format="dmy">17/12/2005</say-as>
<say-as interpret-as="date" format="ydm">2005/17/12</say-as>
<say-as interpret-as="date" format="my">12/2005</say-as>
<say-as interpret-as="date" format="md">12/17</say-as>
<say-as interpret-as="date" format="ym">2005/12</say-as>
</speak>
```

ordinal

This value will speak the ordinal value for the given digit within the enclosed tag.

Example (This will say ″second first″):

```
<speak xml:lang="En-US" version="1.0">
<say-as interpret-as="ordinal">2</say-as>
<say-as interpret-as="ordinal">1</say-as>
</speak>
```

cardinal

This value will speak the cardinal number corresponding to the Roman numeral within the enclosed tag.

Example (This will say ″Super Bowl thirty-nine″):

```
<speak xml:lang="En-US" version="1.0">
Super Bowl <say-as interpret-as="cardinal">XXXIX</say-as>
</speak>
```

number

This value is an alternative to using the values given above. Using the **format** attribute to determine how the number is to be interpreted, you can enter one series of number and have it pronounced several different ways, as in the example. The example also includes two different ways of pronouncing a series of numbers as a telephone number. To have the series pronounced with the punctuation included, you must add the **detail** attribute.

Example:

```
<speak xml:lang="En-US" version="1.0">
<say-as interpret-as="number">123456</say-as>
<say-as interpret-as="number" format="ordinal">123456</say-as>
<say-as interpret-as="number" format="cardinal">123456</say-as>
<say-as interpret-as="number" format="telephone">555-555-5555</say-as>
<say-as interpret-as="number" format="telephone" detail="punctuation">555-555-5555</s
</speak>
```

vxml:boolean

This value will speak ″yes″ or ″no″ depending on the value given within the enclosed tag.

Example:

```
<speak xml:lang="En-US" version="1.0">
<say-as interpret-as="vxml:boolean">true</say-as>
<say-as interpret-as="vxml:boolean">false</say-as>
</speak>
```

vxml:date

This value works like the **date** value, except that the format is predefined as YYYYMMDD. When a value is not known, or you do not wish it to be displayed, a question mark is used to replace that value, as shown in the example.

Example:

```
<speak xml:lang="En-US" version="1.0">
<say-as interpret-as="vxml:date">20050720</say-as>
<say-as interpret-as="vxml:date">????0720</say-as>
<say-as interpret-as="vxml:date">200507??</say-as>
</speak>
```

vxml:currency

This value is used to control the synthesis of monetary quantities. The string must be written in the ″UUUmm.nn″ format, where ″UUU″ is the three character currency indicator specified by ISO standard 4217, and ″mm.nn″ is the amount.

Example (This will say ″forty-five dollars and thirty cents″):

```
<speak xml:lang="En-US" version="1.0">
<say-as interpret-as="vxml:currency">USD45.30</say-as>
</speak>
```

If there are more than two decimal places in the number within the enclosed tag, the amount will be synthesized as a decimal number followed by the currency indicator. If the three character currency indicator is not present, the number will be synthesized as a decimal only, with no pronunciation of currency type.

Example 2 (This will say ″forty-five point three two nine US dollars″):

```
<speak xml:lang="En-US" version="1.0">
<say-as interpret-as="vxml:currency">USD45.329</say-as>
</speak>
```

vxml:phone
This value will speak a phone number with both digits and punctuation, similar to the **number** value used with **format=″telephone″**.

Example:

```
<speak xml:lang="En-US" version="1.0">
<say-as interpret-as="vxml:phone">555-555-5555</say-as>
</speak>
```

<phoneme>
The SSML phoneme tag enables users to provide a phonetic pronunciation for the enclosed text. This tag has two attributes:

alphabet
This attribute specifies the phonology used. The supported alphabets to designate are ″ipa,″ for the International Phonetic Alphabet, and ″ibm,″ for the SPR representation discussed in “Introduction to symbolic phonetic representations” on page 11. The alphabet attribute is optional. If no alphabet is designated, the default value used is ″ibm.″

ph This attribute specifies the pronunciation. It is a required attribute.

This example shows how a pronunciation for ″tomato″ is specified using the IPA phonology, where the symbols are given using Unicode:

```
<speak xml:lang="En-US" version="1.0">
<phoneme alphabet="ipa" ph="t&#x259;mei&#x27E;ou&#x325;">tomato</phoneme>
</speak>
```

This example shows how a pronunciation for ″tomato″ is specified using the SPR phonology:

```
<speak xml:lang="En-US" version="1.0">
<phoneme alphabet="ibm" ph=".0tx.1me.0fo">tomato</phoneme>
</speak>
```

<sub> This tag is used to indicate that the text included in the **alias** attribute is to replace the text enclosed within the tag when speech is synthesized. The only attribute for this tag is the **alias** attribute, and it is required. Without the **alias** attribute defined an error will result.

Example:

```
<speak xml:lang="En-US" version="1.0">
<sub alias="International Business Machines">IBM</sub>
</speak>
```

<voice>

This tag is used when a change in voice is required. Although all attributes listed are optional, without any attributes defined an error will result. The optional attributes are:

age Accepted values are positive integers between the ages of 14 and 60 for both male and female.

gender

Accepted values are ″male″ and ″female″.

name Accepted values are the installed voices' names.

variant

Accepted values are positive integers.

Examples:

```
<speak xml:lang="En-US" version="1.0">
<voice age="any positive integer between 14 and 60">Female voice .</voice>
<voice gender="female">This is a female voice.</voice>
<voice name="Allison">Use the IBM TTS voice named Allison.</voice>
<voice name="Allison, Andrew, Tyler">Use the first available IBM TTS voice named in
</speak>
```

When using voice variant, you must have two female voices of the same language installed.

```
<voice variant="1">Hello, my name is Tyler, I am the
second US English female voice for TTS.</voice>
```

You do not need to specify a voice variant to use the default voice, but to change to a different voice, you must specify the voice variant as ″1″.

<emphasis>

The <emphasis> element is currently not supported.

<break>

This tag inserts pauses into the spoken text. It has the following optional attributes:

strength

This attribute specifies the length of a pause in terms of varying strength values: ″none,″ ″x-weak,″ ″weak,″ ″medium,″ ″strong,″ or ″x-strong.″

time This attribute specifies the length of the pause in terms of seconds or milliseconds. The values formats are ″NNNs″ for seconds or ″NNNms″ for milliseconds.

Example:

```
<speak xml:lang="En-US" version="1.0">
Different sized <break strength="none">pauses.</break>
Different sized <break strength="x-weak">pauses.</break>
Different sized <break strength="weak">pauses.</break>
Different sized <break strength="medium">pauses.</break>
Different sized <break strength="strong">pauses.</break>
Different sized <break strength="x-strong">pauses.</break>
Different sized <break time="1s">pauses.</break>
Different sized <break time="1000ms">pauses.</break>
</speak>
```

<prosody>

This tag controls the pitch, range, speaking rate, and volume of the text.

All attributes are optional, but if no attribute is given an error results. Here is a description of the optional attributes:

pitch This attribute modifies the baseline pitch for the text enclosed within the tag. Accepted values are either:

- a number followed by the Hz designation
- a relative change
- "x-low"
- "low"
- "medium"
- "high"
- "x-high"
- "default"

range This attribute modifies the pitch range for the text enclosed within the tag. Accepted values for this attribute are the same as the accepted values for pitch.

rate This attribute indicates a change in the speaking rate for contained text. Accepted values are:

- a relative change
- a positive number
- "x-slow"
- "slow"
- "medium"
- "fast"
- "x-fast"
- "default"

The rate is specified in terms of words-per-minute. If the speaking rate is 50 words per minute, then rate=50. If the setting is rate=+10, the speaking rate will be 10 words per minute faster than your current rate setting.

Note: When rate is set to a positive number, the implementation is not compliant with the current W3C prosody rate attribute specification.

volume

This attribute modifies the volume for the contained text. The range for values is ″0.0″ to ″100.0″ or the relative values of :

- "silent"
- "x-soft"
- "soft"
- "medium"
- "loud"
- "x-loud"
- "default"

Examples:

```
<speak xml:lang="En-US" version="1.0">
<prosody pitch="150Hz"> Modified pitch </prosody>
<prosody pitch="-20Hz"> Modified pitch </prosody>
<prosody pitch="+20Hz"> Modified pitch </prosody>
```

```
<prosody pitch="-12st"> Modified pitch </prosody>
<prosody pitch="+12st"> Modified pitch </prosody>
<prosody pitch="x-low"> Modified pitch </prosody>
<prosody range="150Hz"> Modified pitch range</prosody>
<prosody range="-20Hz"> Modified pitch range</prosody>
<prosody range="+20Hz">Modified pitch range</prosody>
<prosody range="-12st">Modified pitch range</prosody>
<prosody range="+12st">Modified pitch range</prosody>
<prosody range="x-high">Modified pitch range</prosody>
<prosody rate="slow">Modified speaking rate</prosody>
<prosody rate="+25">Modified speaking rate</prosody>
<prosody rate="-25">Modified speaking rate</prosody>
<prosody volume="88.9">Modified volume</prosody>
<prosody volume="loud">Modified volume</prosody>
</speak>
```

<audio>

This tag inserts recorded elements into the TTS generated audio. The only attribute is **src** and is required. This attribute specifies the location of the file to be inserted.

Example:

```
This is an example of the <audio src="http://www.myfiles.com/files/beep.wav"/> audio being i
```

<mark>

This empty element tag allows the user to place a marker into the text to be synthesized. The synthesis engine notifies the calling program when the engine reaches the marker during synthesis. The mark tag does not affect speech output. It has one required attribute: **name**. The **name** attribute is of the type xsd:token.

Example:

```
<speak xml:lang="En-US" version="1.0">
Example using <mark name="here"/> mark tags.</speak>
```

<lexicon>

This tag introduces pronunciation dictionaries for the given SSML document. The lexicon tag is an immediate child of the speak tag. Its required attribute is **uri**, which specifies the location of the lexicon file.

Example:

```
<speak xml:lang="En-US" version="1.0">
<lexicon uri="http://www.myfiles.com/lexicons.lex"/>
</speak>
```

SSML tips

Spacing and the <sub> element

The following syntax will manifest a problem when spoken:

Example: `<s> The distance is 17<sub alias = ″feet″> ft. </sub></s>`

In this example the TTS engine is supposed to read the word feet normally. Instead, because the `<sub>` element is adjacent to the numeral 17, the word feet is erroneously spelled character-by-character. To resolve, insert a space on either side of the annotation.

Additional information regarding the <sub> element and spaces:

When using the <sub> element, ensure any spaces you need are on the outside of the tag as opposed to inside the tag. Any spaces inside the tag will be replaced by whatever values are in the alias attribute.

For example:

```
This is 3<sub alias="feet"> ft</sub> -----> will become: This is 3feet
This is 3 <sub alias="feet">ft</sub> -----> will become: This is 3 feet
```

Similarly with spaces after the </sub> :

```
This is 3<sub alias="feet"> ft </sub>2 inches -----> will become: This is
3feet2inches
This is 3 <sub alias="feet">ft</sub> 2 inches -----> will become: This is 3
feet 2 in
```

Introduction to symbolic phonetic representations

IBM Text-to-Speech technology uses symbolic phonetic representation (SPR) to represent the sounds of words. This topic briefly defines SPR and provides links to language-specific SPR topics.

SPR is a phonetic coding used to represent the pronunciation of words. An SPR represents the sounds of a word, how the sounds are divided into syllables, and which syllables receive stress.

Symbolic phonetic representation forms (SPRs)

This topic describes the form of a symbolic phonetic representation (SPR).

An SPR consists of a sequence of allowable SPR symbols for a given language, enclosed in quotations and placed within the phoneme tag. For example, the following are valid SPRs in English:

- though `<phoneme alphabet="ibm" ph=".1Tru"> through </phoneme>`
- shocking `<phoneme alphabet="ibm" ph=".1Sa.0kIG"> shocking </phoneme>`

A period signals the beginning of a new syllable, the digits 1 and 0 indicate the stress level of the syllables, and the letters **T**, **r**, **u**, **S**, **a**, **k**, **I**, and **G** represent specific sounds of U.S. English speech. Each of these elements is discussed in more detail in this topic.

Note: An SPR entry that does not conform to the coding requirements is invalid, and is spelled out character by character.

Syllable boundaries

A period is used to mark the beginning of each syllable in the speech generated by the Text-to-Speech technology. However, periods are optional in SPR input in all languages, and, except in German, do not affect how the Text-to-Speech rules divide a word into syllables. by the text-to-speech rules.

In German, a period can be used in SPR input to trigger a syllable boundary at the specified location (see German symbolic phonetic representations).

Syllable stress

Syllables can be marked for stress using the digits 1, or 2, or 0, for primary stress, secondary stress, and no stress, respectively. Some languages do not use secondary stress and thus do not accept the use of the digit 2 in SPRs; see sections on specific languages. If a word has more than one syllable, at least one of these syllables must be marked for primary stress, or the SPR is considered invalid and is read out character by character. Other syllables can be marked with either secondary or no stress. Syllables that are not marked for stress are assumed to have no stress.

The syllable stress marker (1, 2, or 0) should be within a syllable boundary, but always to the left of the vowel of the syllable. The marker can be placed anywhere to the left of the stressed syllable, as shown in the following example.

Suppose you do not know where the syllable boundaries are located in the word **construction**. In this example, any of the following SPRs correctly place the primary stress on the highlighted vowel:

"**construction**"
"kXn1str**H**kSXn"
"kXns1tr**H**kSXn"
"kXnst1r**H**kSXn"
"kXnstr1**H**kSXn"

Speech sound symbols

Each language uses its own inventory of SPR symbols for representing the speech sounds of that language. Tables in the following sections contain valid SPR symbols for the sounds of each language, with examples of words in which each sound occurs. Letters are case-sensitive, so "e" and "E", for example, represent two different sounds. Two-character symbols must be contained in single quotes; for example, German **heim** "h'aj'm". SPRs containing sound symbols that are not allowed in a current language are considered invalid, and are spelled out character by character.

The sounds of every language have specific distributional patterns within that language. For example, in all dialects of English, the sound "G" in **sing** ".1sIG" does not occur at the beginning of a word. Other American English sounds that have a particularly narrow distribution are the glottal stop "?", the flap "F", and the syllabic nasal "N". If you enter a sound symbol in a context where it does not normally occur, the resulting speech may sound unnatural.

IBM Text-to-Speech technology applies a sophisticated set of linguistic rules to its input to reflect the processes by which sounds change in specific contexts in natural language. For example, in American English, the sound "t" of **write** ".1r1Yt" is pronounced as a flap "F" in **writer** ".1rY.0FR". SPR input undergo these modifications just as ordinary input text does. In this example, whether you enter ".1rY.0tR" or ".1rY.0FR" does not affect the generated speech.

Entering symbolic phonetic representations

You can enter symbolic phonetic representations (SPRs) to replace the pronunciation of a word.

U.S. English SPRs

This topic contains tables that list the U.S. English symbolic phonetic representations (SPRs) by category.

The following tables show the inventory of valid SPR symbols in U.S. English. Each sound symbol is accompanied by examples showing typical spellings of the sound in actual words, with the letters representing the given sound underlined. (Because of dialectal differences, the examples might not always match your pronunciation.)

Regular vowels

The following table lists the symbols for the regular vowels.

U.S. English symbol	Example words
a	father, lot
A	back, had
e	cake, pain
E	hedge, let
i	see, speak, believe
I	pick, ill
o	both, oak
c	law, cough
u	zoo, truth
U	took, put
H	but, mug, son
R	butter, hurt

Diphthongs

The following table lists the symbols for diphthongs.

U.S. English symbol	Example words
O	toil, boy
W	out, cow
Y	life, fine

Reduced vowels

The following table lists the symbols for reduced vowels.

U.S. English symbol	Example words
x	sofa, alone, suppose, tedious, America
X	roses, connect, melody, symphony, hinted

Consonants

The following table lists the symbols for the consonants.

U.S. English symbol	Example words
b	bad, sob
p	pit, rip
d	dip, had
t	tip, pet
g	good, bug
k	kill, cat, make, back
D	this, breathe
T	thing, Beth
v	vase, save
f	field, if, graph
z	zip, phase
s	seal, miss, ceiling
Z	treasure, garage
S	ship, wish
J	Jane, huge
C	chip, witch, nature
h	hot, hero
m	man, hum, summer
n	never, sun, winner
G	sing, finger
r	borrow, rake
l	low, hall
w	wear, quick
y	yes, Virginia
M	hmmm
? (″glottal stop″)	kitten, Latin
F (″flap″)	writer, fiddle
N (″syllabic nasal″)	button, satin, eaten, burden

Syllable stress

The following table lists the symbols for syllable stress

1	Primary stress (most prominent stress in the word)
2	Secondary stress
0	No stress

Syllable boundary

The following table lists the symbol for syllable boundary.

. (period)	Beginning of a syllable

Related information

"Symbolic phonetic representation forms (SPRs)" on page 11
This topic describes the form of a symbolic phonetic representation (SPR).

"Entering symbolic phonetic representations" on page 12
You can enter symbolic phonetic representations (SPRs) to replace the pronunciation of a word.

British English SPRs

This topic contains tables that list the British English symbolic phonetic representations (SPRs) by category.

The following tables show the inventory of valid SPR symbols in British English. Each sound symbol is accompanied by examples showing typical spellings of the sound in actual words, with the letters representing the given sound underlined. (Because of dialectal differences, the examples might not always match your pronunciation.)

Regular vowels

The following table lists the symbols for the regular vowels.

British English symbol	Example Words
a	path, father, chant
A	back, had
e	cake, pain
E	hedge, let
i	see, speak, believe
I	pick, ill
o	both, oak
c	law, court, hall, water
@	rod, cough
u	zoo, truth
U	took, put
H	but, mug, son
R	butter, hurt

Diphthongs

The following table lists the symbols for diphthongs.

British English symbol	Example words
O	toil, boy
W	out, cow

British English symbol	Example words
Y	life, fine

Reduced vowels

The following table lists the symbols for reduced vowels.

British English symbol	Example words
x	sofa, alone, suppose, America
X	roses, hinted

Consonants

The following table lists the symbols for the consonants.

British English symbol	Example Words
b	bad, sob
p	pit, rip
d	dip, had
t	tip, pet
g	good, bug
k	kill, make, back
D	this, breathe
T	thing, Beth
v	vase, save
f	field, if, graph
z	zip, phase
s	seal, miss, ceiling
Z	treasure, garage
S	ship, wish
J	Jane, huge
C	chip, witch, nature
h	hot, hero
m	man, hum, summer
n	never, sun, winner
G	sing, finger
r	borrow, rake
l	low, hall
L	candle
w	wear, quick
y	yes, Virginia

Syllable stress

The following table lists the symbols for syllable stress

1	Primary stress (most prominent stress in the word)
2	Secondary stress
0	No stress

Syllable boundary

The following table lists the symbol for syllable boundary.

. (period)	Beginning of a syllable

Related information

"Symbolic phonetic representation forms (SPRs)" on page 11
This topic describes the form of a symbolic phonetic representation (SPR).

"Entering symbolic phonetic representations" on page 12
You can enter symbolic phonetic representations (SPRs) to replace the pronunciation of a word.

German SPRs

This topic contains tables that list the U.S. English symbolic phonetic representations (SPRs) by category.

The following tables show the inventory of valid SPR symbols in German. Each sound symbol is accompanied by examples showing typical spellings of the sound in actual words, with the letters representing the given sound underlined. (Because of dialectal differences, the examples might not always match your pronunciation.)

Remarks specific to SPRs in German are provided in the appropriate sections.

Vowels

German symbol	Example words
i	lieben, Titel, tief
I	bitte, Tisch, Licht
e	geben, Ehre, See
E	treffen, Geld, kämmen
'E:'	Käse, Mädchen, wägen
a	Haar, haben, fahren
A	lassen, matt, Apfel
u	gut, Uhr, Uwe
U	Hund, Fluß, Mutter
o	Ober, ohne, Boot
O	Kopf, Stopp
y	Bücher, Tür, kühn
Y	fünf, füllen, Künstler
'oe'	Löwe, hören, Söhne

German symbol	Example words
'OE'	können, hölzern, östlich
@	bitte, Kamera, Boden

Diphthongs

The following table lists the symbols for diphthongs.

German symbol	Example words
'aj'	heim, Waise, Mai
'aw'	Haus, Maul, Frau
'oj'	heute, Gebäude, Häuser

Nasalized Vowels

The following table includes the symbols for nasalized vowels.

German symbol	Example words
'a~'	Chance
'E~'	Teint
'o~'	Pardon
'oe~'	Parfum

Consonants

The following table lists the symbols for the consonants.

German symbol	Example words
b	Boden, Bett, oben
p	Papier, Lippe, Grab
d	dunkel, kindisch, Helden
t	Tag, bitte, Rad
g	geben, grau, Tage
k	Katze, Ecke, Skulptur, lag, quitt
v	Wagen, viskös, Volum, oval
f	fast, hoffen, Vater
z	See, Satz, lesen
s	Fuß, lassen, Last, Haus
Z	Garage, Genie
S	schon, spielen, Stil, wäscht
X	ich, Chemie, Kelch, mancher
x	Buch, Bach, Wochen
P	Pflanze, Stumphen
T	Zauber, Polizei, Glanz
J	Job, Dschungel

German symbol	Example words
C	deutsch, Chile, Cello
m	Mann, kommen, Atem
n	Nacht, können, Kind
G	Finger, längs, Anfang
l	lesen, fallen, Pult
r	Rad, führen
R	Wieder, über
j	Junge, ja, Jahr, Ministerium
w	Eduard, aktuell, Januar
h	hoch, Hand, Ahorn

Syllable stress

The following table lists the symbols for syllable stress

1	Primary stress (most prominent stress in the word)
2	Secondary stress
0	No stress

Syllable boundary

Note: In German, a period in an SPR entry triggers a syllable boundary at that location.

The following table lists the symbol for syllable boundary.

. (period)	Beginning of a syllable

Related information

"Symbolic phonetic representation forms (SPRs)" on page 11
This topic describes the form of a symbolic phonetic representation (SPR).

"Entering symbolic phonetic representations" on page 12
You can enter symbolic phonetic representations (SPRs) to replace the pronunciation of a word.

Canadian French SPRs

This topic contains tables that list the Canadian French symbolic phonetic representations (SPRs) by category.

The following tables show the inventory of valid SPR symbols in Canadian French. Each sound symbol is accompanied by examples showing typical spellings of the sound in actual words, with the letters representing the given sound underlined. (Because of dialectal differences, the examples might not always match your pronunciation.)

Remarks specific to SPRs in Canadian French are provided in the appropriate sections.

Vowels

The following table lists the symbols for the vowels.

French symbol	Example words
a	pattes, lac, cave
A	char, mâle
e	café, déformer, été
E	faite, dresser
i	film, typique
I	site, plastique, ride
o	taurillon, vaudevilliste
c	paul, note, échalotte
u	roue, où, tour
U	foule, mousse
y	utile, pure, Bruno
Y	autobus, chute
x	litres, marbre (note: le [x] s'efface dans certains contextes.)
'eu'	meuglement
'oe'	cependant, cheval
'a:'	voyage, information
'e:'	steak (anglicismes)
'E:'	père, annuaire, fête
'o:'	paule, beau, tôt, côté
'c:'	loge, encore
'u:'	four, douze
'y:'	dur, buse
'eu:'	jeûne, émeute
'oe:'	peur, jeune, déjeuner
'a~'	banc, en, temps
'E~'	fin, plein, faim
'o~'	bon, pont, mon
'oe~'	un, aucun

Consonants

The following table lists the symbols for the consonants.

French symbol	Example words
b	bébé, balle, robe
p	porte, prêt, guêpe
d	dort, dolmen
t	ton, patte, théâtre
g	guuerre, bague, garer

French symbol	Example words
k	kilo, caler, quai
v	laver, wagon, visiter
f	chef, faim, phare
D	duque, dire
T	petit, tuque
z	jaser, réseau, zigzaguer
s	sans, ambition, façon
Z	rage, gîte, jouer
S	cheval, lâche, schéma
J	jeans, jogging
C	gaucho, gaspa**ch**o
m	maman, femme, miser
n	Anne, ni, maniaque
′nj′	agneau, campagne
′ng′	parking, camping
r	parer, rare, carreau
l	litre, illisible, pâle
j	hiérarchie, paille, yoga
w	oui, bouée, watt
H	sua, lui, nuée

Syllable stress

The following table lists the symbols for syllable stress

1	Primary stress (most prominent stress in the word)
2	Secondary stress
0	No stress

Syllable boundary

The following table lists the symbol for syllable boundary.

. (period)	Beginning of a syllable

Liaison

In French, the underscore can be used following a word-final consonant (but within the right bracket which closes the SPR) to indicate that it is a liaison consonant: that is, it will be pronounced only if the following word begins with a vowel.

For example, a roots dictionary key *petit* with the translation value ″p′oe′t1it_″ will have the final ″t″ pronounced in the input string *un petit ami* but not in the input

string *un petit chien*. On the other hand, an entry with the translation value "nEt" will have the final "t" pronounced regardless of context.

The following examples show how to use the symbol for liaison.

_	(underscore character) allow liaison if the following word begins with a vowel. For example:	
	"p0'oe't1it_"	The "t" is not pronounced unless the following word begins with a vowel.
	"nEt"	The "t" is always pronounced.

Related information

"Symbolic phonetic representation forms (SPRs)" on page 11
This topic describes the form of a symbolic phonetic representation (SPR).

"Entering symbolic phonetic representations" on page 12
You can enter symbolic phonetic representations (SPRs) to replace the pronunciation of a word.

Spanish SPRs

This topic contains tables that list the symbolic phonetic representations (SPRs) for Mexican and Castilian Spanish, with differences between the two in notes as needed.

The following tables show the inventory of valid SPR symbols in Mexican and Castilian Spanish. Each sound symbol is accompanied by examples showing typical spellings of the sound in actual words, with the letters representing the given sound underlined. (Because of dialectal differences, the examples might not always match your pronunciation.)

Vowels

The following table lists the symbols for the regular vowels.

Spanish symbol	Example words
a	agua
e	este
i	igual
o	oso
u	uva

Consonants

The following table lists the symbols for the consonants.

Spanish symbol	Example words
b	basta, beber, vaca
p	parte, apagar
d	dar, dedo
t	toma, atar
g	goma, haga

Spanish symbol	Example words
k	cuenco
f	flaco, afuera
z	mismo, desde
s	silla, casa
R	ropa, perro
T	zapato, socio
C	chalupa, mucho
j	junco, reja, gente
m	mano, amor
n	nada, mano
N	piña
r	para, pero
l	loco, algo
L	llover, pollo
Y	yegua, playa
y	medio, oigo
w	fuera, deuda

Note: The phonetic symbol T is realized only in Castilian Spanish, and is replaced internally with phonetic symbol s in Mexican Spanish, even when present in the SPR input.

The following table lists the internal allophones (for information only).

Allophone symbol	Example word
B (b)	robo
D (d)	hada
G (g)	segar
'ng' (n)	ancla, engaño

Syllable stress

The following table lists the symbols for syllable stress

1	Primary stress (most prominent stress in the word)
2	Secondary stress
0	No stress

Syllable boundary

The following table lists the symbol for syllable boundary.

. (period)	Beginning of a syllable

Related information

"Symbolic phonetic representation forms (SPRs)" on page 11
This topic describes the form of a symbolic phonetic representation (SPR).

"Entering symbolic phonetic representations" on page 12
You can enter symbolic phonetic representations (SPRs) to replace the pronunciation of a word.

Introduction to pitch and its use with SSML

This section is about pitch fluctuations, and how pitch is used within the context of IBM TTS and SSML.

What is pitch?

The terms *pitch* and *pitch range* are familiar to musicians. Pitch is usually specified as the name of a note and an octave number. For example, A4 is the note "A" in octave 4. Pitch range is specified as the number of octaves that an instrument or a singer can cover, from the lowest note to the highest. For example, if the lowest note is A2 and the highest is A4, then the range is two octaves.

In the so-called tempered scale, each octave is divided into 12 semitones.

Each note corresponds to sound vibrations at a particular frequency, measured in Hertz (Hz), or cycles per second. For example the frequency of the note A4 is 440 Hz. An interval of one octave corresponds to a doubling or halving of the frequency. Thus A3 is at 220 Hz, and A5 is at 880 Hz.

The frequency of any note can be calculated from the formula $f = 440 \cdot 2^{n/12}$ where f is the frequency in Hz and n is the number of semitones between A4 and the note in question. Thus, A4# is one semitone higher, so its frequency is $440 \cdot 2^{1/12} = 466.16376$ **Hz**. A5 is an octave higher, that is 12 semitones, so its frequency is $440 \cdot 2^{12/12} = 440 \cdot 2 = 880$ **Hz**.

What is meant by the pitch range of a speaking voice?

Although the pitch range of a keyboard instrument is well defined, the pitch range of a speaking voice is not. Because of phenomena such as glottalization, it is possible for the time interval between pitch periods to occasionally be very large, which, if taken literally, would imply an extremely low pitch frequency. Similarly, there may be occasional excursions to high pitches not really characteristic of the speaker's general style. To achieve a more stable and meaningful measure of the speaker's pitch range, we define the 5-th percentile as the bottom of the range and the 95-th percentile as the top. In other words, the pitch range is defined so that the speaker's pitch stays in that range 90% of time. During 5% of the time the pitch is actually below the "bottom" and during another 5% it is above the "top." These extreme excursions are considered outliers, and not part of the normal pitch range.

The outliers, outside the normal pitch range, occur infrequently enough that in many single-sentence utterances, they may not occur at all. In fact, most short utterances will have a much narrower pitch range than the specified nominal range. Thus, if you request a pitch range of 200 Hz, and ask the synthesizer to say "Hello," you would not expect this short utterance to cover the full 200-Hz range.

What is base pitch?

The term*base pitch* can be defined as the average frequency of the speaker's voice, measured in Hz. This means that the bottom of the pitch range will be below the base pitch, and the top of the range will be above it.

What is the relationship between pitch range and base pitch?

Normally, a higher base pitch goes along with a higher pitch range, when measured in Hz. If you raise the base pitch, but the pitch range in Hz is not changed, the voice begins to sound monotone. If the range is measured in semitones, however, it need not be changed when the base pitch is changed.

The larger the pitch range, the greater the difference between the lowest pitch and the base pitch, and the same goes for the difference between the base and the highest pitch. In other words, as the pitch range increases, the bottom pitch drops and the top pitch increases.

If I ask for a base pitch of 100 Hz and a range of 200 Hz, will the bottom of the range be 0 Hz and the top 200 Hz?

No. The top of the pitch range will be about 230 Hz, and the bottom will be about 30 Hz.

The pitch range is not centered at base pitch, when measured in Hz. The difference between the base pitch and the lowest pitch will be smaller than the difference between the base pitch and the highest pitch. In this example, the bottom is 70 Hz below the base, but the top is 130 Hz above the base pitch. The frequency of the bottom pitch is constrained to be non-negative, but there is no mathematical constraint on the top of the pitch range.

In technical terms—it is assumed the pitch distribution is log-normal in the frequency domain, which leads to a normal Gaussian bell curve in the semitone domain. The mean and the median of the log-normal density function are not equal. If you define the base pitch to be the mean, then median will therefore decrease if the pitch range is increased, while keeping the base pitch constant.

To explain this in regular terms—if you ask for a large pitch range, but specify a low base pitch, then overall pitch will spend most of the time well below the base, or average value, but will have occasional peaks at values much higher than the average so that the average still comes out correctly. For example, if the base pitch is 100 Hz and the range is 200 Hz, then the top of the pitch range will be about 230 Hz, and the bottom will be 30 Hz. Even though the average pitch is 100 Hz, the pitch contour will actually spend more than half of the time below 100 Hz. In fact, half the time it will be below 83 Hz, but during the other half it will go high enough to bring the overall average up to 100 Hz.

Are all combinations of base pitch and range possible?

Although MRCP and SSML allow base pitch and pitch range to be specified independently, not all combinations are usable.

The synthesizer will not produce a pitch lower than 50 Hz. This limit may be encountered if the base pitch is fairly low, for example 100 (typical for male speakers) and the requested pitch range is large, for example, 200 Hz. In that case, the bottom of the pitch range would be about 30 Hz, which is below the capability of the synthesizer. If you want such a large pitch range, you may want to experiment with slightly higher base pitches.

In general, pitch ranges that are equal to or greater than double the base pitch, in Hz, may cause difficulty. In semitones, this means a range of greater than 36 semitones, or 3 octaves.

As a rule, the default values for base pitch and range will produce the best sound quality, but you may want to change them for specific effects.

TTS question intonation

This topic details how to write questions so that the TTS engine pronounces them with the proper intonation.

To avoid ambiguous intonations from the TTS engine when developing your application, write questions using a yes-no or wh-type question format. Wh-type questions ask for who-what-when-where and why information, and are the most common in the English language. The yes-no type format asks direct questions requiring only an affirmative or negative response from the user. These types of questions like, ″What is your PIN?″ and ″Do you want to cancel the transaction?″ are understood and read with the appropriate intonation by the TTS engine.

However, some yes-no questions differ from a declarative sentence only by intonation. For example, the rising pitch at the end of ″You requested a fund transfer?″ indicates that it is a question and not the declarative statement, ″You requested a fund transfer.″ Reword this sentence and replace it with the less ambiguous, ″Did you request a fund transfer?″ to avoid questions being read as declarative statements by the TTS engine.

Notices and trademarks

The following paragraph does not apply to the United Kingdom or any country where such provisions are inconsistent with local law:

INTERNATIONAL BUSINESS MACHINES CORPORATION PROVIDES THIS PUBLICATION "AS IS" WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF NON-INFRINGEMENT, MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. Some states do not allow disclaimer of express or implied warranties in certain transactions; therefore, this statement may not apply to you.

This information could include technical inaccuracies or typographical errors. Changes are periodically made to the information herein; these changes will be incorporated in new editions of the publication. IBM may make improvements and/or changes in the product(s) and/or the program(s) described in this publication at any time without notice.

Any references in this information to non-IBM Web sites are provided for convenience only and do not in any manner serve as an endorsement of those Web sites. The materials at those Web sites are not part of the materials for this IBM product and use of those Web sites is at your own risk.

IBM may use or distribute any of the information you supply in any way it believes appropriate without incurring any obligation to you.

Licensees of this program who wish to have information about it for the purpose of enabling: (i) the exchange of information between independently created programs and other programs (including this one) and (ii) the mutual use of the information which has been exchanged, should contact:

IBM Corporation
TL3B/B503
3039 Cornwallis Road
Research Triangle Park, NC 27709
U.S.A

Such information may be available, subject to appropriate terms and conditions, including in some cases, payment of a fee.

The licensed program described in this document and all licensed material available for it are provided by IBM under terms of the IBM Customer Agreement, IBM International Program License Agreement or any equivalent agreement between us.

Any performance data contained herein was determined in a controlled environment. Therefore, the results obtained in other operating environments may vary significantly. Some measurements may have been made on development-level systems and there is no guarantee that these measurements will be the same on generally available systems. Furthermore, some measurement may have been estimated through extrapolation. Actual results may vary. Users of this document should verify the applicable data for their specific environment.

Information concerning non-IBM products was obtained from the suppliers of those products, their published announcements or other publicly available sources. IBM has not tested those products and cannot confirm the accuracy of

performance, compatibility or any other claims related to non-IBM products. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products.

All statements regarding IBM's future direction or intent are subject to change or withdrawal without notice, and represent goals and objectives only.

COPYRIGHT LICENSE:

This information contains sample application programs in source language, which illustrates programming techniques on various operating platforms. You may copy, modify, and distribute these sample programs in any form without payment to IBM, for the purposes of developing, using, marketing or distributing application programs conforming to the application programming interface for the operating platform for which the sample programs are written. These examples have not been thoroughly tested under all conditions. IBM, therefore, cannot guarantee or imply reliability, serviceability, or function of these programs. You may copy, modify, and distribute these sample programs in any form without payment to IBM for the purposes of developing, using, marketing, or distributing application programs conforming to IBM's application programming interfaces.

Each copy or any portion of these sample programs or any derivative work, must include a copyright notice as follows: (C) (your company name) (year). Portions of this code are derived from IBM Corp. Sample Programs. (C) Copyright IBM Corp. 2005. All rights reserved. If you are viewing this information softcopy, the photographs and color illustrations may not appear.

Trademarks

The following terms are trademarks or registered trademarks of the International Business Machines Corporation in the United States, other countries, or both:

IBM

Other company, product, and service names may be trademarks or service marks of others.

Index

Readers' Comments — We'd Like to Hear from You

IBM Text-to-Speech SSML Programming Guide

We appreciate your comments about this publication. Please comment on specific errors or omissions, accuracy, organization, subject matter, or completeness of this book. The comments you send should pertain to only the information in this manual or product and the way in which the information is presented.

For technical questions and information about products and prices, please contact your IBM branch office, your IBM business partner, or your authorized remarketer.

When you send comments to IBM, you grant IBM a nonexclusive right to use or distribute your comments in any way it believes appropriate without incurring any obligation to you. IBM or any other organizations will only use the personal information that you supply to contact you about the issues that you state on this form.

Comments:

Thank you for your support.

Submit your comments using one of these channels:

- Send your comments to the address on the reverse side of this form.
- Send a fax to the following number: 1-800-227-5088 (US and Canada)

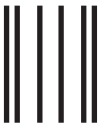
If you would like a response from IBM, please fill in the following information:

Name	Address
Company or Organization	
Phone No.	E-mail address

Readers' Comments — We'd Like to Hear from You

Cut or Fold
Along Line

Fold and Tape | **Please do not staple** | Fold and Tape



NO POSTAGE
NECESSARY
IF MAILED IN THE
UNITED STATES

BUSINESS REPLY MAIL

FIRST-CLASS MAIL PERMIT NO. 40 ARMONK, NEW YORK

POSTAGE WILL BE PAID BY ADDRESSEE

IBM Corporation
Information Development
Department MMOA
P.O. Box 12195
Research Triangle Park, NC 27709-9990

Fold and Tape | **Please do not staple** | Fold and Tape

Cut or Fold
Along Line