VOXYGEN

VOXYGEN | SSML

Reference manual

Core version: 9.4

Reference	:	VOX31_SSML_reference_manual_9.4_1.0_EN			
Doc. version	:	1.0	**Status**	:	Release
Date	:	03/05/2024	**Disclosure**	:	Restricted

Contents

1 PRESENTATION 3
1.1 About this document 3
1.2 Terminology 3
1.3 Reference documents 3

2 TEXTUAL INPUT 3

3 OUTPUT 4
3.1 Errors 4
3.2 Synchronization 4

4 LANGUAGES AND VOICES 4
4.1 Installed voices 4
4.2 Voice selection 5
4.3 xml:lang attribute values 6

5 SSML ELEMENTS 7
<audio> 8
<vox:audiomix> 10
<break> 11
<vox:checksum> 12
<desc> 13
<emphasis> 13
<lang> (SSML 1.1) 14
<lexicon> 14
<lookup> (SSML 1.1) 15
<mark> 16
<meta> 17
<metadata> 17
<p> 18
<phoneme> 19
<prosody> 19
<s> 23
<say-as> 24
<speak> 25
<sub> 26
<vox:token>/<vox:w> (SSML 1.0) <token>/<w> (SSML 1.1) 27
<vox:version> 28
<voice> 28

1 Presentation

1.1 About this document

This document describes issues specific to the implementation of SSML in Voxygen Core. The current implementation covers both SSML version 1.0 as described in the W3C Recommendation 7 September 2004: http://www.w3.org/TR/speech-synthesis/ and SSML version 1.1 as described in the W3C Recommendation 7 September 2010: http://www.w3.org/TR/speech-synthesis11/.

In this document, the word 'platform' refers to a piece of software in which Voxygen Core is integrated. The availability of some SSML features in Voxygen Core is dependent on the way the TTS system is integrated into the platform. Such cases are identified below. Some additional elements and attributes are implemented in Voxygen Core as extensions to the W3C recommendations for SSML. These extensions are described herein.

1.2 Terminology

Abbreviation	Description
SSML	Speech Synthesis Markup Language
PLS	Pronunciation Lexicon Specification
TTS	Text To Speech
Baratinoo	Code name for the Voxygen Core (Voxygen TTS engine)

1.3 Reference documents

Reference	Document name
VOX170	Exceptions lexicon reference manual
VOX341	Text Normaliser reference manual
VOX349	Phonemes and visemes reference manual
VOX370	Data reference manual

2 Textual input

XML documents are assumed to be encoded in **Unicode-UTF8** unless the encoding attribute in the xml prologue specifies otherwise. It is recommended to always specify a document's encoding scheme.

Supported values of the **encoding** attribute are (case-insensitive): "UTF-8", "UTF-16", "UTF-16BE", "UTF-16LE", "ISO-8859-1" and "US-ASCII".

3 Output

3.1 Errors

The XML processor included in the TTS system is conforming and validating. The way an error is reported depends on the platform. Baratinoo starts to process an SSML document before it is totally fetched. Thus a non well-formed document may be partially processed (speech signal produced) before an error is detected.

3.2 Synchronization

Baratinoo may also report synchronization events to the platform. The following events may be reported:

- a change in voice, which may be triggered by the <voice> element or a change in the value of xml:lang,
- marker, triggered by a <mark> element (of type "sync").

Events are reported to the platform when they arrive at the output of the system. For example, a voice event occurs immediately before the first sample of speech of the new voice is given to the platform.

4 Languages and voices

Languages supported by Baratinoo are generally not all installed onto the platform. Furthermore, for a given installed language, often not all of the voices the system can synthesize are installed either. It is thus necessary to know what is installed and how the system selects the current voice in order to obtain the best results from the system.

4.1 Installed voices

Please refer to the main Voxygen product documentation to know how to install the distribution and configure the list of voices. In short, the file baratinoo.cfg in the directory <VOXYGEN_MAIN_DIRECTORY>/data/9.4/ may contain the list of voices. This list can be edited; the order of the voices and the voice names can be changed.

Each voice has the following attributes:

- name
- gender
- language + optional region
- version
- speech modes

The first voice in the list (voice number 0) is considered as the “default voice”, unless a different voice is first selected via the Baratinoo API (which is platform dependent).

4.2 Voice selection

Voice selection criteria depend on information provided by the values of the xml:lang attribute – which may be given in the <speak>, <p>, <s>, <token>, <lang> (SSML 1.1) or <voice> (SSML 1.0 only) elements – and by attributes name, gender, age, variant and languages (SSML 1.1) of the <voice> element. The order of importance and necessity for a voice to match requested voice selection criteria are controlled in SSML 1.1 by the values of the required and ordering attributes.

The algorithm used to select a voice is as follows:

1) A list of voices is initialized from the order in which they are specified in the Baratinoo configuration file.

2) All available voices are identified for which the values of all voice feature attributes listed in the required attribute value are matched. When the value of the required attribute is the empty string "", any and all voices are considered successful matches. If one or more voices are identified, the selection is considered successful; otherwise there is voice selection failure.

3) If a successful selection identifies only one voice, the synthesis processor uses that voice.

4) If a successful selection identifies more than one voice, the remaining features (those not listed in the required attribute value) are used to choose a voice by feature priority (see 6), where the starting candidate set is the set of all voices identified.

5) To choose a voice by feature priority, each feature is taken in turn starting with the highest priority feature, as controlled by the ordering attribute.

   - If at least one voice matches the value of the current voice feature attribute then all voices not matching that value are removed from the candidate set. If a single voice remains in the candidate set the synthesis processor uses it. If more than one voice remains in the candidate set then the next priority feature is examined for the candidate set.
   - If no voices match the value of the current voice feature attribute then the next priority feature is examined for the candidate set.

6) After examining all feature attributes on the ordering list, if multiple voices remain in the candidate set, then other features (not yet considered), excluding the variant and vox:version features, are given equal priority and the voices that satisfy the most features are kept in the candidate set. The variant and vox:version features are then used in an attempt to further reduce the set of candidates. If a single voice remains in the candidate set the synthesis processor uses it.

7) If more than one voice remains in the candidate set the synthesis processor removes candidates that cannot speak the current language given by xml:lang. If no candidates match, it restores the set. If a single voice remains in the candidate set the synthesis processor uses it.

8) Finally, if the existing voice is in the set of remaining candidates, the processor keeps that voice; otherwise it uses the first variant from the list of remaining candidates.

The Baratinoo voice selection algorithm for SSML 1.0 documents is equivalent to specifying required="name" ordering="languages name gender age variant" and languages = current xml:lang value.

All Baratinoo voices can read text written in the language they speak wherever the text was written. For example, an English voice that speaks with a British accent can read text written in both American and British styles of English.

A British voice may be selected to speak English text by using the attribute languages="en:en-GB" in SSML 1.1. Note that in SSML 1.0 there is no mechanism to select a voice's regional accent.

Despite Baratinoo voices being monolingual (an onlangfailure will occur when given text from another language), they may be able to use xml:lang information to assist them with reading foreign words if the onlangfailure attribute is set to "ignorelang" (the xml:lang value is not totally ignored, it's just ignored as far as voice selection is concerned).

Example : A list of installed voices:

1	Marion	female	French (fr:fr-FR)
2	Arnaud_neutre	male	French (fr:fr-FR)
3	Jenny	female	English, US accent (en:en-US)
4	Paul	male	English, GB accent (en:en-GB)

And a sample SSML document:

```
<speak version="1.0"
       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-synthesis/synthesis.xsd
```

```
            http://www.w3.org/2005/01/pronunciation-lexicon
            http://www.w3.org/TR/pronunciation-lexicon/pls.xsd"
      xml:lang="fr-FR">
  <!-- no hints => default voice i.e. Marion, so default xml:lang is "fr" -->
  Bonjour, je suis Marion.
  <voice gender="male">
    <!-- {"fr" implicit, "male"} => Arnaud_neutre -->
    Bonjour, je suis Arnaud.
    <p xml:lang="en">
    <!-- change voice onlangfailure {"en", "male" implicit} => Paul -->
      Hello, I am Paul.
      <voice gender="female">
        <!-- {"en" implicit, "female"} => Jenny -->
        Hello, I am Jenny.
        <s xml:lang="en-GB">
          <!-- no onlangfailure, so no voice change{"en-GB","female"}=>Jenny -->
          Hello, I am still Jenny.
        </s>
      </voice>
    </p>
  </voice>
</speak>
```

4.3 xml:lang attribute values

The valid values of the xml:lang attribute are IETF BCP47 language tags: http://www.ietf.org/rfc/bcp/bcp47.txt. Baratinoo accepts full BCP47 syntax but uses only the language code, and an optional region code. The simplified syntax is: language[-region]. Baratinoo accepts all values but not all values are supported.

The values supported for the language code (ISO 639) that may have an effect on rendering cover French (fr, fre), German (de, deu, ger), English (en, eng), Spanish (es, spa), Italian (it, ita), Russian (ru, rus), Arabic (ar, ara), Polish (pl, pol), Latin (la, lat), Basque (eu, eus, baq), Swedish (sv, swe), Dutch (nl, nld, dut), Hungarian (hu, hun), Romanian (ro, ron, rum) and Japanese (ja, jpn).

The values supported for the region code (ISO 3166-1 or UN M.49) that may have an effect on rendering cover Belgium (BE, BEL, 056), Canada (CA, CAN, 124), Switzerland (CH, CHE, 756), France (FR, FRA, FX, FXX, 250), Great-Britain (GB, GBR, UK, 826), Ireland (IE, IRL, 372), and the United-States (US, USA, 840).

Although Baratinoo is case-insensitive for language and region codes, it is recommended to respect the lower-case of ISO 639 language codes and upper-case of ISO 3166-1 region codes.

5 SSML elements

Here is a list of supported SSML elements and their attributes. See the W3C SSML recommendations for more details.

The second column ("St") of the attribute tables indicates the status of each attribute:

- **M:** mandatory (must/required; error if not present)
- **R:** recommended (should; warning if not present)
- **O:** optional (may; ok if not present)
- **E:** Baratinoo extension (optional)

Baratinoo provides some proprietary extensions to the W3C recommendations for SSML:

- Elements: <token>/<w> (SSML 1.0); <audiomix>, <version>, and <checksum>.
- Attributes:
  - diacritics and pauses for <speak>, <p> and <s>;

- modes for <speak>, <p>, <s> and <w>;
- version for <voice>;
- timbre, computedpitch and computedduration for <prosody>;
- gain, fadein, fadeout and fadelevel for <audio>;
- idl for <phoneme>; and
- type for <mark>.
- Relaxed constraints: the name attribute of the <mark> element is recommend, rather than required.
- Functionalities: interaction between <say-as> and a user lexicon.

A specific namespace has been defined for extension elements, attributes and QName values: http://www.voxygen.fr/tts

This document suggests using vox as the namespace prefix. Please refer to the XML namespace recommendation for details: http://www.w3.org/TR/REC-xml-names/.

When strict parsing is disabled in the Baratinoo .cfg file, the extension namespace may be omitted.

Any XML elements or attributes that are encountered from a declared namespace other than those defined within SSML and our extension namespaces will be ignored.

<audio>

Description

Insert a recorded audio file.

If the audio file cannot be retrieved, the element's contents are synthesized.

Attribute	St	Value
src	M (SSML 1.0) O (SSML 1.1)	Name of file (absolute or relative URI)
fetchtimeout (SSML 1.1)	O	Signed or unsigned positive number or zero followed by "s" for seconds or "ms" or milliseconds. Default "30s".
fetchhint (SSML 1.1)	O	"prefetch" (default) "safe"
maxage (SSML 1.1)	O	A positive integer or zero.
maxstale (SSML 1.1)	O	A positive integer or zero.
clipBegin (SSML 1.1)	O	Signed or unsigned positive number or zero followed by "s" for seconds or "ms" or milliseconds. Default "0s".
clipEnd (SSML 1.1)	O	Signed or unsigned positive number or zero followed by "s" for seconds or "ms" or milliseconds.
repeatCount (SSML 1.1)	O	Signed or unsigned positive number or zero. Default "1".
repeatDur (SSML 1.1)	O	Signed or unsigned positive number or zero followed by "s" for seconds or "ms" or milliseconds.
soundLevel (SSML 1.1)	O	Signed number followed by "dB" for decibels. Default "+0.0dB".
speed (SSML 1.1)	O	Unsigned positive number or zero followed by "%". Default "100%".

Possible content
audio, vox:audiomix, break, vox:checksum, emphasis, lang (SSML 1.1), lookup (SSML 1.1), mark, phoneme, prosody, say-as, voice, sub, p, s, vox:token/vox:w (SSML 1.0), token/w (SSML 1.1), and desc elements. Text.

Restrictions

On some platforms, an HTTP/FTP client is embedded in Baratinoo to download audio files from a server. URI schemes other than file:, http: and ftp: in the value of the src attribute (after application of xml:base) are not supported.

The MIME type is determined from the resource file suffix and introspection rather than from any response in the transfer protocol.

file suffix	MIME type
.au	audio/x-au
.wav	audio/x-wav
.a8k .alaw	audio/x-alaw-basic
.raw .ulaw	audio/basic

Only audio/x-wav files may be in mono or stereo. Other media types must contain only one signal channel (mono).

The speed attribute is truncated to the interval [50%;200%].

The soundLevel attribute is truncated to the interval [-90.0dB;+12dB].

If the repeatDur or repeatCount attribute is used, the maximum duration of the audio insertion is 5 minutes.

Baratinoo extension

Baratinoo introduces four new attributes to the <audio> element that modify the audio source.

Attribute	St	Value
vox:gain	E	Signed number followed by "dB" for decibels. Default "+0.0dB".
vox:fadelevel	E	Signed number followed by "dB" for decibels. Default "+0.0dB".
vox:fadein	E	Signed or unsigned positive number or zero followed by "s" for seconds or "ms" or milliseconds. Default "0s".
vox:fadeout	E	Signed or unsigned positive number or zero followed by "s" for seconds or "ms" or milliseconds. Default "0s".
vox:fadeinAttack	E	Positive number or zero followed by "s" for seconds or "ms" or milliseconds. Default "20ms".
vox:fadeinRelease	E	Positive number or zero followed by "s" for seconds or "ms" or milliseconds. Default "20ms".
vox:fadeoutAttack	E	Positive number or zero followed by "s" for seconds or "ms" or milliseconds. Default "20ms".
vox:fadeoutRelease	E	Positive number or zero followed by "s" for seconds or "ms" or milliseconds. Default "20ms".
vox:tempo	E	Unsigned positive number followed by "%". Default "100%".

The gain attribute is provided for SSML 1.0 as an alias to the SSML 1.1 soundLevel attribute. In SSML 1.1, if both soundLevel and gain attributes are given, soundLevel takes precedence.

The fadein specifies a time relative to the beginning of the element. The fadeout specifies a time relative to the end of the element. The soundLevel (alias gain) attribute specifies the relative volume of the referenced audio from the time specified by the fadein attribute to the time specified by the fadeout attribute. The volume specified by soundLevel overrides the volume specified by the value of the fadelevel attribute during this period. The fadelevel

attribute specifies the relative volume outside this period. In order to avoid audible artefacts, the sound level *may* change progressively during short transitional periods (typically 20ms) at the beginning of the fadein and at the end of the fadeout.

If the fadein instance occurs after the fadeout instance, then the relative volume of the entire referenced audio is specified by the fadelevel attribute.

The soundLevel and fadelevel attributes are truncated to the interval [-90.0dB;+12dB].

The fadein and fadeout attributes are truncated to the interval [0s;60s].

The default duration (20ms) of the transitional periods at either side of the fadein and the fadeout periods may be changed. The fadeinAttack and fadeinRelease attributes are respectively for the beginning side and the end side of the fadein period. The same with the fadeoutAttack and fadeoutRelease attributes for the fadeout period. The sum of the attack and release value for a period must not be greater than the duration of the period.

The tempo attribute can be used to speed up or slow down the rate of the audio file without changing the pitch level.

The tempo attribute is truncated to the interval [50%;200%].

<vox:audiomix>

Description

Mix a recorded audio file with the element content.

If the audio signal is longer than the synthesis of the element's content, then the audio signal is truncated. If the audio signal is shorter than the synthesis of the element's content, then the synthesis processor repeatedly reads the file. If the audio file cannot be retrieved, only the element's content is synthesized.

Attribute	St	Value
src	M	Name of file (absolute or relative URI)
fetchtimeout	O	Signed or unsigned positive number or zero followed by "s" for seconds or "ms" or milliseconds. Default "30s".
fetchhint	O	"prefetch" (default) "safe"
maxage	O	A positive integer or zero.
maxstale	O	A positive integer or zero.
clipBegin	O	Signed or unsigned positive number or zero followed by "s" for seconds or "ms" or milliseconds. Default "0s".
clipEnd	O	Signed or unsigned positive number or zero followed by "s" for seconds or "ms" or milliseconds.
soundLevel	O	Signed number followed by "dB" for decibels. Default "+0.0dB".
speed	O	Unsigned positive number or zero followed by "%". Default "100%".
gain	O	Signed number followed by "dB" for decibels. Default "+0.0dB".
fadelevel	O	Signed number followed by "dB" for decibels. Default "+0.0dB".
tempo	O	Unsigned positive number followed by "%". Default "100%".
fadein	O	Signed or unsigned positive number or zero followed by "s" for seconds or "ms" or milliseconds. Default "0s".
fadeout	O	Signed or unsigned positive number or zero followed by "s" for seconds or "ms" or milliseconds. Default "0s".
fadeinAttack	O	Positive number or zero followed by "s" for seconds or "ms" or milliseconds. Default "480ms".

Attribute	St	Value
fadeinRelease	O	Positive number or zero followed by "s" for seconds or "ms" or milliseconds. Default "480ms".
fadeoutAttack	O	Positive number or zero followed by "s" for seconds or "ms" or milliseconds. Default "480ms".
fadeoutRelease	O	Positive number or zero followed by "s" for seconds or "ms" or milliseconds. Default "480ms".

Possible content
audio, vox:audiomix, break, vox:checksum, emphasis, lang (SSML 1.1), lookup (SSML 1.1), mark, phoneme, prosody, say-as, voice, sub, p, s, vox:token/vox:w (SSML 1.0), token/w (SSML 1.1), and desc elements. Text.

Note

Attributes of the <audiomix> element have the same meaning and restrictions as those of the <audio> element, but the default fade attack and release durations may differ.

The values of attributes affect only the audio source prior to mixing with the element content.

<break>

Description

Empty element that controls the pausing or prosodic boundaries between words.

Attribute	St	Value
time	O	Signed or unsigned positive number or zero followed by "s" for seconds or "ms" or milliseconds. Extension: percentage values are also accepted.
strength	O	"none" ≅ 0ms "x-weak"≅ 50ms "weak" ≅ 100ms "medium" ≅ 500ms (default) "strong" ≅ 1s "x-strong" ≅ 2s

Possible content
Nothing

Note

If both time and strength attributes are supplied, the system uses a break with a duration as specified by the `time` attribute.

Use of the strength attribute may have an effect on the prosody as well as the break duration.

Restrictions

Break duration has an upper limit of 60 seconds. This element may be placed in succession should a longer break duration be required.

Baratinoo extension

Inside a <prosody> element with a duration attribute, <break> may have a time attribute whose value is a percentage. If the text is longer than the required duration, any % break will be removed. If the required duration of the prosody is longer than the text inside, effective duration of the <break> whose time is a percentage will be computed. In the following example, if the text duration is 1s, the first break duration will be 1.2s (30% of 4s) and the second break duration will be 2.8s (70% of 4s).

Example:

```
<prosody duration="5s">Time for<break time="30%"/>a pause<break
time="70%"/></prosody>
```

Furthermore, the duration of each break will be at most (% of the break * prosody_duration). In the following example, even if the text is short, the break will be only 1s long (and so the text will be spoken slowly to fill in the remaining 9s)

Example:

```
<prosody duration="10s">Short<break time="10%"/>text</prosody>
```

A break time of "0%" is interpreted as "0ms".

<vox:checksum>

Description

Enable a cyclic-redundancy check to be preformed on the signal and events in the most recent breath group (delimited by silence) rendered from the content of the current document.

Attribute	St	Value
crc32	M	unsigned positive integer or zero

Possible content
Nothing

<desc>

Description

Inside an audio element only, describes the content of the audio source.

Attribute	St	Value
xml:lang	O	see xml:lang section
onlangfailure (SSML 1.1)	O	"changevoice" "ignoretext" "ignorelang" "processorchoice" (Default value is inherited from parent)

Possible content
Text that is to be displayed in text-only mode.

Restrictions

This element and its contents are ignored by Baratinoo since it does not provide text-only output.

<emphasis>

Description

Requests that the contained text be spoken with emphasis.

Attribute	St	Value
level	0	"none" "reduced" "moderate" (default) "strong"

Possible content
audio, vox:audiomix, break, vox:checksum, emphasis, lang (SSML 1.1), lookup (SSML 1.1), mark, phoneme, prosody, say-as, voice, sub, vox:token/vox:w (SSML 1.0), and token/w (SSML 1.1) elements. Text.

Restrictions

The realization of emphasis is voice dependent.

<lang> (SSML 1.1)

Description

Specify the natural language of the written content.

Attribute	St	Value
xml:lang	M	See xml:lang section
onlangfailure	0	"changevoice" "ignoretext" "ignorelang" "processorchoice" (Default value is inherited from parent)

Possible content
audio, vox:audiomix, break, vox:checksum, emphasis, lang (SSML 1.1), lookup (SSML 1.1), mark, phoneme, prosody, say-as, voice, sub, p, s, vox:token/vox:w (SSML 1.0), and token/w (SSML 1.1) elements. Text

<lexicon>

Description

Defines a user lexicon to be used when processing text in the SSML document. More than one <lexicon> element

can be used in an SSML document, each one defining a user lexicon.

Attribute	St	Value
uri	M	Location of the lexicon document.
xml:id (SSML 1.1)	M	A unique identifier for the lexicon document.
type	O	Preferred media type of the lexicon document.
fetchtimeout (SSML 1.1)	O	Signed or unsigned positive number or zero followed by "s" for seconds or "ms" or milliseconds. Default "30s".
maxage (SSML 1.1)	O	A positive integer or zero.
maxstale (SSML 1.1)	O	A positive integer or zero.

Possible content
Nothing

Note

The <meta>, <metadata> and <lexicon> elements must occur before all other elements and text contained within the root <speak> element.

In SSML 1.0, the referenced lexicon is immediately active and remains so throughout the scope of the current <speak> element. In SSML 1.1, however, the referenced lexicon is inactive until it is explicitly activated by a <lookup> element.

When multiple lexicons are active in SSML 1.0, their precedence goes from lower to higher with document order. Precedence means that a token is first looked up in the lexicon with highest precedence. Only if not found in that lexicon, is the next lexicon searched and so on until a first match or until all lexicons have been used for lookup.

In SSML 1.1, lexicon lookup precedence is governed by the <lookup> element.

A lexicon document is typically used to define pronunciations externally. In contrast to pronunciations specified by a <phoneme> element, a token's pronunciation issued from a lexicon may be subject to contextual modifications for the realisation of liaison and/or assimilation, since lexical pronunciations are defined context-free.

Restrictions

On some platforms, an HTTP/FTP client is embedded in Baratinoo to download lexicons from a server. URI schemes other than file:, http: and ftp: in the value of the uri attribute (after application of xml:base) are not supported.

Lexicon files can be in one of two formats:

- XML file, conformant with PLS 1.0 as described by the W3C Recommendation 14 October 2008: http://www.w3.org/TR/pronunciation-lexicon/.
- Plain text file, using a proprietary syntax.

See document "VOX170 Exceptions lexicon reference manual" for more details.

The media type of the lexicon document is always determined by introspection rather than through using a transfer protocol. The value of the type attribute is not used.

Baratinoo extension

Baratinoo allows a proprietary lexicon file format:

- text-preprocessing RGX file. This file must be a conformant XML document.

See document "VOX341 Text normaliser reference manual" for more details.

<lookup> (SSML 1.1)

Description

Activate a lexicon so that the information it contains is used by the synthesis processor when rendering tokens that appear within the content of the element.

Attribute	St	Value
ref	M	Specifies a name that references a lexicon as assigned by the `xml:id` attribute of the `<lexicon>` element

Possible content
audio, vox:audiomix, break, vox:checksum, emphasis, lang (SSML 1.1), lookup (SSML 1.1), mark, phoneme, prosody, say-as, voice, sub, p, s, vox:token/vox:w (SSML 1.0), and token/w (SSML 1.1) elements. Text

<mark>

Description

Places a marker into the text/tag sequence.

Attribute	St	Value
name	R (Baratinoo) M (SSML)	Marker name

Possible content
Nothing

Baratinoo extension

Baratinoo introduces are new attribute that allows a marker to be designated as a *wait* marker, as opposed to the default SSML sync marker.

Attribute	St	Value
vox:type	E	`"sync"` (default) "wait"

A wait marker allows rendering of the audio signal to be deferred until the duration of the immediately following content has been determined. The end of the content whose duration is to be determined is marked by either the end of the root <speak> element or a <mark> element, of any type, that bears the same name (case-sensitive).

For example:

```
“Text before…<mark name="foo" vox:type="wait"/> piece of text <mark name="foo"/> text after…”
```

When Baratinoo processes the above markup, notification is first made by a 'WAITMARKER' event with the name 'foo' and the duration in samples of the rendered content "piece of text". Then the signal for the "piece of text" is sent, and finally, notification is made by a 'MARKER' event with the name 'foo', signaling the end of the marked sequence.

It is possible to set another <mark>, of any type, before the end of the deferred content is encountered. Examples are:

```
…<mark name="foo" vox:type="wait"/> piece of text <mark name="another"/> containing another marker <mark name="foo"/>
```

```
…<mark name="foo" vox:type="wait"/> piece of text <mark name="another"
vox:type="wait"/> with another embedded <mark name="another"/> wait marker
sequence<mark name="foo"/>
…<mark name="foo" vox:type="wait"/> piece of text <mark name="another"
vox:type="wait"/> with interleaved <mark name="foo"/> wait marker sequences<mark
name="another"/>
```

The name attribute is not required by Baratinoo (but it is recommended). If omitted, the marker name takes the value of an integer counter, and the type of the marker is forced to the default value (i.e. the type attribute value is ignore). The counter starts at 1 for the current session and is incremented by one each time an unnamed <mark> element is encountered in a document. The automatically attributed name is sent in the ensuing 'MARKER' event.

Markers whose names are attributed automatically are never allowed to match the markers referenced by the startmark and endmark attributes of the <speak> element, or by the name of wait makers.

<meta>

Description

Empty container by which information about the document can be provided.

Attribute	St	Value
http-equiv name	M	A property name
content	M	Value to be associated with the named property

Possible content
Nothing

Note

The <meta>, <metadata> and <lexicon> elements must occur before all other elements and text contained within the root <speak> element.

Restrictions

Baratinoo processes only the HTTP property named "Content-Location". If xml:base is not already attributed by the <speak> element, then the associated value must have correct IETF RFC3986 syntax and it will be used by Baratinoo as the base URI; otherwise the value is ignored.

<metadata>

Description

Container in which information about the document can be placed, using a metadata schema.

Possible content
XML data (the content is ignored)

Note

The <meta>, <metadata> and <lexicon> elements must occur before all other elements and text contained within the root <speak> element.

The contents of this element may be used cunningly to hold elements in the PLS namespace, which can describe pronunciations required when reading the current SSML document. The <lexicon> element uri attribute can then be used to reference the current document.

For example, an SSML 1.1 document called "ssml+pls.xml" may contain:

```
<lexicon uri="./ssml+pls.xml" xml:id="this"/>
<metadata>
          <pls:lexicon version="1.0" alphabet="ipa" xml:lang="en">
          <pls:lexeme>   …   </pls:lexeme>
          </pls:lexicon>
</metadata>
```

Restrictions

This element and its contents are ignored when the document is parsed as an SSML document.

<p>

Description

Identifies the enclosed text as a paragraph.

Attribute	St	Value
xml:lang	O	See xml:lang section
onlangfailure (SSML 1.1)	O	"changevoice" "ignoretext" "ignorelang" "processorchoice" (Default value is inherited from parent)
xml:id (SSML 1.1)	O	A unique identifier for the current element

Possible content
audio, vox:audiomix, break, vox:checksum, emphasis, lang (SSML 1.1), lookup (SSML 1.1), mark, phoneme, prosody, say-as, voice, sub, s, vox:token/vox:w (SSML 1.0), and token/w (SSML 1.1) elements. Text.

Restrictions

This element does not replace paragraph chunking; it just marks the beginning and the end of a paragraph.

Baratinoo extension

Baratinoo introduces new attributes to the <p> element that provide control over the way the text content is processed. Their function is the same as those defined in the <speak> element.

Attribute	St	Value
vox:diacritics	E	"default" "non" "acc" (Default value is inherited from parent)
vox:pauses	E	"default" "syntagma" "punctuation" (Default value is inherited from parent)
vox:modes	E	A space-separated list of speech mode names.

<phoneme>

Description

Provides a phonetic pronunciation for the contained text.

Attribute	St	Value
ph	M	List of phonetic symbols (separated by underscore '_' when x-voxygen alphabet is used). See VOX349 document for specifications of phonetic alphabets.
alphabet	O	"x-voxygen" (default) "ipa"
type (SSML 1.1)	O	"default" (default) "ruby"

Possible content
Text.

Note

The provided pronunciation is considered to be definitive for the contained text. It is thus not subject to contextual modifications for the realization of liaison and/or assimilation.

Restrictions

Support of the IPA alphabet is limited to sounds that map to the phonetic symbols of the current voice.

The value of the ph attribute is ignored for unsupported alphabets and a warning is issued.

Baratinoo extension

Baratinoo introduces a new attribute to control the inclusion or exclusion of specific acoustic units as candidate realisations for each part of the given phonetic pronunciation.

Attribute	St	Value
vox:idl	E	[*ids*] *pho* [*ids*] *pho*... [*ids*] *ids* = list of comma separated acoustic unit identifiers (integers). An identifier may be preceded by + for 'inclusion' during unit selection, otherwise 'exclusion' from unit selection is inferred. *Pho* = a x-voxygen phonetic symbol

<prosody>

Description

Permits control of the pitch, speaking rate and volume of the speech output.

Attribute	St	Value
pitch	O	"x-low" = 50% of "default" "low" = 75% of "default" "medium" = 100% of "default" "high" = 133% of "default" "x-high" = 200% of "default" "default" = initial value for current voice
		Relative percentage ([+/-] number followed by "%")

Attribute	St	Value	
		Relative change ([+/-] number followed by "Hz" for Hertz or "st" for semitones)	
		Absolute value in Hertz (unsigned number followed by "Hz")	
range	0	"x-low" = 50% of "default" "low" = 75% of "default" "medium" = 100% of "default" "high" = 133% of "default" "x-high" = 200% of "default" "default" = initial value for current voice	
		Relative percentage ([+/-] number followed by "%")	
		Relative change ([+/-] number followed by "Hz" for Hertz or "st" for semitones)	
		Absolute value in Hertz (unsigned number followed by "Hz")	
rate	0	"x-slow" = 50% of "default" "slow" = 75% of "default" "medium" = 100% of "default" "fast" = 125% of "default" "x-fast" = 150% of "default" "default" = initial value for current voice	
		Relative percentage ([+/-] number followed by "%") *Extension of SSML 1.1.*	
		Relative change ([+/-] number with no units) *Extension of SSML 1.0 & SSML 1.1.*	
		Absolute value: multiplier of the "default" rate (unsigned number with no units in SSML 1.0; followed by "%" in SSML 1.1). Extension: "%" is optional in Baratinoo. *Extension: rate factor may affect only speech or pauses (see vox:rate-subject)*	
volume	0	"silent" "x-soft" relative to "default" "soft" relative to "default" "medium" relative to "default" "loud" relative to "default" "x-loud" relative to "default" "default" = initial value for current voice	
		Relative percentage ([+/-] number followed by "%") on linear scale. *Extension of SSML 1.1*	
		Relative change ([+/-] number with no units) on linear scale. *Extension of SSML 1.1*	
		Relative change in dB ([+/-] number followed by "dB") *Extension of SSML 1.0*	
		Absolute value in interval [0;100], linear scale. *Extension of SSML 1.1.*	
duration	0	Signed or unsigned positive number or zero followed by "s" for seconds or "ms" or milliseconds.	
contour	0	List of pair of the form (time position, target)	
		time position	percentage of period (number followed by "%") in interval [0%;100%]
		target	a pitch attribute value

Possible content
audio, vox:audiomix, break, vox:checksum, emphasis, lang (SSML 1.1), lookup (SSML 1.1), mark, phoneme, prosody, say-as, voice, sub, p, s, vox:token/vox:w (SSML 1.0), and token/w (SSML 1.1) elements. Text.

Note

Relative percentages (i.e. signed numbers followed by "%") of a prosodic feature (pitch, range and volume) are multipliers of its *current* value, which may have already been changed by a parent <prosody> element. Relative changes also apply to the *current* value.

new value = current value × (1 + attribute value)

Absolute percentages (i.e. unsigned numbers followed by "%") of a prosodic feature are multipliers of the voice's *initial* value for the prosodic feature. Labeled values are also modifiers of the voice's *initial* value.

new value = initial value × attribute value

When a voice change occurs within a prosody element, any absolute or accumulated relative changes made to prosodic features are imposed on the new voice relative to its initial values.

Restrictions

The effects of prosodic attribute values are limited. The absolute value of a prosodic feature after having applied the required modification is truncated to reside within the following ranges:

Attribute	Minimum	Maximum
pitch	30Hz	–
range	0Hz	300Hz
pitch+range	–	600Hz
rate	×0.1	×10
volume	-90dB	+24dB
`contour(target)`	20Hz	450Hz

Unsigned relative percentages are allowed values for the pitch, range, and volume attributes and target. This is conformant with SSML 1.0 and extended to SSML 1.1. A warning is issued if this depreciated format is used. An unsigned percentage for the rate attribute in SSML 1.0 is not interpreted as a relative change. It is interpreted as an absolute percentage of the voice's initial speaking rate (as in SSML 1.1).

Baratinoo internally uses dB for values of volume. In Baratinoo, the default value of the volume attribute is +0.0dB (as specified in SSML 1.1). SSML 1.0 documentation specifies the default value as 100, but Baratinoo may not conform with this since it does not necessarily correspond to +0.0dB. There is a transformation from the SSML 1.0 linear scale [0;100] to decibels, which may be controlled in the system's configuration file. The minimal value, 0, is silent and the maximal value, 100, depends on the configuration. The normal configuration is:

Attribute Value	SSML Value	dB
silent	0	-∞dB
x-soft	20	-12dB
soft	40	-6dB
medium	`60` (default)	0dB
loud	80	+6dB
x-loud	100	+12dB

If a contour attribute has been specified, child <prosody> elements are ignored. If a duration and rate attributes are specified, rate is applied first, then duration, so as to give priority to the duration attribute. A <prosody> element with

a rate attribute is not allowed to contain a child <prosody> element with a duration attribute, but the inverse is allowed.

Baratinoo extension

Baratinoo introduces new attributes to the <prosody> element that allow voice timbre to be altered, make rate affect only speech or pauses, and to enable/disable prosody computation.

Attribute	St	Value
vox:timbre	E	Relative percentage ([+/-] number followed by "%")
		Relative value ([+/-] number with no units)
		Absolute value: multiplier of the initial timbre value for the current voice (unsigned number with no units or followed by "%").
vox:rate-subject	E	"articulation": rate value affects only speech "pause": rate value affects only pauses originated from the synthesis engine (<break> value are not affected) "all": rate value affects both speech and pauses (default value)
vox:computedpitch	E	"on": apply pitch contour computed by system "off": intrinsic pitch contour "default": reset to default behaviour of voice
vox:computedduration	E	"on": apply phoneme duration computed by system "off": intrinsic phoneme duration "default": reset to default behaviour of voice

The timbre attribute is a rate/pitch warping coefficient that maintains the duration of phonemes and enables voice timbre to be modified.

The rate-subject attribute indicates what shall be affected by rate factor : speech, pauses or both. If unspecified, rate applies to both.

When a <prosody> element requires a pitch modification (via the contour attribute), the computed pitch mode is automatically switched to "on" and is reset to the previous value at the end of the processing, unless the computed mode is explicitly set to "off" using the vox:computedpitch attribute. In which case, the pitch modification may not work as expected.

<s>

Description

Identifies the enclosed text as a sentence.

Attribute	St	Value
xml:lang	O	See xml:lang section
onlangfailure (SSML 1.1)	O	"changevoice" "ignoretext" "ignorelang" "processorchoice" (Default value is inherited from parent)
xml:id (SSML 1.1)	O	A unique identifier for the current element

Possible content
audio, vox:audiomix, break, vox:checksum, emphasis, lang (SSML 1.1), lookup (SSML 1.1), mark, phoneme, prosody, say-as, voice, sub, vox:token/vox:w (SSML 1.0), and token/w (SSML 1.1) elements. Text.

Restrictions

This element does not replace sentence chunking; it just marks the beginning and the end of a sentence.

Baratinoo extension

Baratinoo introduces new attributes to the <s> element that provide control over the way the text content is processed. Their function is the same as those defined in the <speak> element.

Attribute	St	Value
vox:diacritics	E	"default" "non" "acc" (Default value is inherited from parent)
vox:pauses	E	"default" "syntagma" "punctuation" (Default value is inherited from parent)
vox:modes	E	A space-separated list of speech mode names.

<say-as>

Description

Indicates the type of text construct contained within the element, as described in the W3C Note 25 May 2005 on say-as attribute values: http://www.w3.org/TR/ssml-sayas/.

Attribute	St	Value
interpret-as	M	A QName. "date" "time" "telephone" "characters" "cardinal" "ordinal"
format	O	string
detail	O	string

Possible content
Text.

Restrictions

The attribute values that may have an effect on rendering depend on the current voice.

The detail attribute value is not used.

Baratinoo extension

If the value of the interpret-as attribute is in Baratinoo's extension namespace, then it may interact with lookups in active lexicons. If a token from the element's content is found in an active lexicon with its vox:say-as attribute equal to the current interpret-as value (once stripped of its namespace), the token is pronounced as given by the lexicon. If a token exists in an active lexicon but without any vox:say-as attribute, the token is also pronounced as given by the lexicon.

<speak>

Description

Root of an SSML document.

Attribute	St	Value
version	M	"1.0" (default) "1.1"
xml:lang	M	See xml:lang section. The default value is provided by the default voice.
onlangfailure (SSML 1.1)	O	"changevoice" "ignoretext" "ignorelang" "processorchoice" (default)
xmlns	M	"http://www.w3.org/2001/10/synthesis"
xmlns:xsi	R	"http://www.w3.org/2001/XMLSchema-instance" This is required if xsi:schemaLocation is to be given.
xsi:schemaLocation	R	"http://www.w3.org/2001/10/synthesis http://www.w3.org/TR/speech-synthesis/synthesis.xsd"
xmlns:vox	O	"http://www.voxygen.fr/tts" This is required only if Baratinoo extensions are used and strict parsing is enabled in the .cfg file.
xml:base	O	Specify the base URI of the root document.
startmark (SSML 1.1)	O	Reference to a marker as assigned by the name attribute of a <mark> element, at which point rendering starts.
endmark (SSML 1.1)	O	Reference to a marker as assigned by the name attribute of a <mark> element, at which point rendering ends.

Possible content
audio, vox:audiomix, break, vox:checksum, emphasis, lang (SSML 1.1), lookup (SSML 1.1), mark, phoneme, prosody, say-as, voice, sub, p, s, vox:token/vox:w (SSML 1.0), token/w (SSML 1.1), lexicon, vox:version, meta and metadata elements. Text.

Note

The default value for the SSML 1.1 onlangfailure attribute, "processorchoice", is chosen to be equivalent to "changevoice" for Baratinoo, so that behaviour is close to that of SSML 1.0. However, this default value can be configured in the baratinoo.cfg configuration file in the [INPUT] section:

```
ssml.onlangfailure.processorchoice = changevoice | ignoretext | ignorelang
```

Baratinoo extension

If strict parsing is disabled in the baratinoo.cfg file then all attributes of this element become optional under SSML 1.0; but their status do not change under SSML 1.1.

Baratinoo introduces new attributes to the <speak> element that provide control over the way the text content is processed.

Attribute	St	Value
vox:diacritics	E	"default" (default) "non" "acc"
vox:pauses	E	"default" (default) "syntagma" "punctuation"
vox:modes	E	A space-separated list of speech mode names.

The diacritics attribute controls text processing with respect to the presence or absence of diacritics. When set to "acc", text is to be processed as if diacritics are present, if any. When set to "non", text is to be processed as if diacritics may have been removed. Set this attribute to "default" to use the system's default behaviour (voice specific) for text processing with respect to the presence or absence of diacritics.

The pauses attribute controls where pauses may be introduced by the system. When set to "syntagma", the system may place pauses wherever it believes there needs to be a strong discontinuity between syntactic groups. When set to "punctuation", the insertion of these possible pauses is reserved to where a punctuation mark is present in the text. Set this attribute to "default" to use the system's default behaviour (voice specific) for the placement of pauses.

The values of the modes attribute provides the names of speech modes which will be active during the selection of acoustic units. The list of available names depends on the voice being used.

<sub>

Description

Text substitution for the purposes of pronunciation.

Attribute	St	Value
alias	M	String to be used in substitution of the element content when it is to be in spoken form.

Possible content
Text.

<vox:token>/<vox:w> (SSML 1.0)
<token>/<w> (SSML 1.1)

Description

In an SSML 1.0 document <vox:token> and <vox:w> are Baratinoo extensions that form aliases for the SSML 1.1 <token> (alias <w>) element.

The <token> element can be used to:

- indicate its content is a token and to eliminate token (word) segmentation ambiguities of the synthesis processor.
- select a specific element in a PLS lexicon in accordance with its role attribute value.
- set a specific grammatical tag for the content text. In this case, the value of role must be a Baratinoo specific value (inside the extension namespace); otherwise it will not be considered as a grammatical tag.

Attribute	St	Value
role	O	A QName used in conjunction with lexicons
`xml:lang` (SSML 1.1)	O	See xml:lang section
`onlangfailure` (SSML 1.1)	O	"changevoice" "ignoretext" "ignorelang" "processorchoice" (Default value is inherited from parent)
`xml:id` (SSML 1.1)	O	A unique identifier for the current element

Possible content
audio, vox:audiomix, break, vox:checksum, emphasis, mark, phoneme, prosody, say-as, and sub elements. Text.

Restrictions

The correct rendering of child elements is not guaranteed when they are not adjacent to the elements markup, since a grapheme-to-phoneme correspondence is required. Elements that break the text content are interpreted as if they were placed at the end of all the text content. When possible, it is recommended to place the <token> element inside the content of other elements.

Baratinoo extension

Baratinoo introduces a new attribute to the <w> element that provide control over the way the text content is processed. Its function is the same as that defined in the <speak> element.

Attribute	St	Value
vox:modes	E	A space-separated list of speech mode names.

<vox:version>

Description

Generates synthesized speech from the name of the TTS engine and its version number.

Possible content
Nothing

<voice>

Description

Specifies voice characteristics for the contained text.

Attribute	St	Value
xml:lang (SSML 1.0 only)	0	See xml:lang section.
gender	0	"male" "female" "neutral" "" (SSML 1.1, initial value)
age	0	Positive integer or zero. "" (SSML 1.1, initial value)
variant	0	Positive integer, not zero. "" (SSML 1.1, initial value)
name	0	Space separated list of voice names. See list of installed voices. "" (SSML 1.1, initial value)
languages (SSML 1.1)	0	List of space-separated languages the voice is desired to speak. "" (initial value)
required (SSML 1.1)	0	A list of space-separated feature names from "gender", "age", "variant", "languages", "name". "" Initial value is "languages"
ordering (SSML 1.1)	0	A list of space-separated feature names from "gender", "age", "variant", "languages", "name". "" Initial value is "languages"
onvoicefailure (SSML 1.1)	0	"priorityselect" (initial value) "keepexisting" "processorchoice"

Possible content
audio, vox:audiomix, break, vox:checksum, emphasis, lang (SSML 1.1), lookup (SSML 1.1), mark, phoneme, prosody, say-as, voice, sub, p, s, vox:token/vox:w (SSML 1.0), and token/w (SSML 1.1) elements. Text.

Note

Default values for omitted parameters are inherited from a parent <voice> element.

For the SSML 1.1 onvoicefailure attribute, the value "processorchoice" is chosen by default to be equivalent to "priorityselect" for Baratinoo. However, this default value can be configured in the baratinoo.cfg configuration file in the [INPUT] section:

```
ssml.onvoicefailure.processorchoice = priorityselect | keepexisting
```

Baratinoo extension

In SSML 1.0, if no attribute is given, a warning is issued and the system switches to the default voice. This is equivalent to specifying required="variant" and variant="1" in SSML 1.1.

The values of the name and gender attributes are case-insensitive.

SSML document content may be tuned for a specific version of a named voice. This is particularly the case when unit selection information is given in the value of the idl attribute of the <phoneme> element.

Attribute	St	Value

vox:version	E	List of space-separated versions the voice is desired to have. "" (initial value)

vox:version is an extension attribute indicating a processor-specific version of a voice to speak the contained text. The value may be a space-separated list of versions ordered from top preference down or the empty string "".

The value of the required and ordering attributes may include a QName. This allows the value to include the name of an attribute in a different namespace, such as "vox:version".

Note that a voice selection failure must be reported by a conforming synthesis processor and that a voice selection failure occurs if no voice can be found that matches all voice feature attributes listed in the required attribute.

For example, if the voice named "Jenny" exists only in version "2.0", but version "2.1" is required with:

```
<voice required="name languages vox:version" name="Jenny" vox:version="2.1">
```

then a voice selection failure is reported.