Language Manual

HQ Indian English

Language Manual: HQ Indian English

Published 16 March 2011
Copyright © 2009-2011 Acapela Group

This document was produced by Acapela Group. We welcome and consider all comments and suggestions. Please, use the *Contact Us* link at our website: http://www.acapela-group.com

Table of Contents

List of Tables

Chapter 1. General

This document discusses certain aspects of text-to-speech processing for the Indian English text-to-speech system, in particular the different types of input characters and text that are allowed.

This version of the document corresponds to High Quality (HQ) voice Deepa.

Please note that the *User's Guide*, mentioned several times in the manual, is called *Help* in some applications.

> Note: For the sake of efficiency, the processing described in this document has a different behaviour in some Acapela Group products. Those products are:
>
> • Acapela TTS for Windows Mobile
>
> • Acapela TTS for Linux Embedded
>
> • Acapela TTS for Symbian
>
> For these products, the default processing of numbers, phone numbers, dates and times has been simplified for the low memory footprint (LF) voice formats. Developers have the possibility to change the default behaviour from *simplified* to *normal* preprocessing by setting corresponding parameters in the configuration file of the voice. Please see the documentation of these products for more information. In the following chapters, each simplification will be described by the indication *[not SP]* following the description of the standard behaviour. The *SP* in the indication stands for *Simplified Processing*.

Chapter 2. Letters in orthographic text

Characters from *A-Z* and *a-z* may constitute a word. Certain other characters are also considered as letters, notably those used as letters in other European languages, i.e. *ñ, ç, é*. These letters are not pronounced as in their native languages though, they are pronounced as regular *n, c, e* when occurring in a word

Characters outside of these ranges, i.e. numbers, punctuation characters and other non-alphanumeric characters, are not considered as letters.

Chapter 3. Punctuation characters

Punctuation marks appearing in a text affect both rhythm and intonation of a sentence. The following punctuation characters are permitted in the normal input text string: *, : ; " " . ? ! ( ) [ ] { }*

3.1. Comma, colon and semicolon

Comma ',', colon ':' and semicolon ';' cause a brief pause to occur in a sentence, accompanied by a small rising intonation pattern just prior to the character.

3.2. Quotation marks

Quotes '""' appearing around a single word or a group of words cause a brief pause before and after the quoted text.

3.3. Full stop

A full stop '.' is a sentence terminal punctuation mark which causes a falling end-of-sentence intonation pattern and is accompanied by a somewhat longer pause. A full stop may also be used as a decimal marker in a number (see chapter *Number processing*) and in abbreviations (see chapter *Abbreviations*).

3.4. Question mark

A question mark '*?*' ends a sentence and causes question-intonation, first rising and then falling.

3.5. Exclamation mark

The exclamation mark '*!*' is treated in a similar manner to the full stop, causing a falling intonation pattern followed by a pause.

3.6. Parentheses

Parenthesis '*()*', brackets '*[]*', and braces '*{}*' appearing around a single word or a group of words cause a brief pause before and after the bracketed text.

Chapter 4. Other non alphanumeric characters

4.1. Non-punctuation characters

The characters listed below are processed as non-letter, non-punctuation characters. Some are pronounced at all times and others are only pronounced in certain contexts, which are described in the following sections of this chapter.

Table 4.1. Non-punctuation characters

Symbol	Reading
/	slash
+	plus
$	dollar
£	pound
€	euro
¥	yen
<	less than
>	greater than
%	percent
^	circumflex
|	pipe
~	tilde
@	at
=	equals
^{2}	(see below)
^{3}	(see below)
-	(see below)
*	(see below)

4.2. The ^{2} and ^{3} signs

The reading of expressions with ^{2} and ^{3} is:

Expression	Reading
mm^2	square millimeters
cm^2	square centimeters
m^2	square meters
km^2	square kilometers
mm^3	cubic millimeters
cm^3	cubic centimeters
m^3	cubic meters
km^3	cubic kilometers

4.3. Symbols whose pronunciation varies depending on the context

4.3.1. Hyphen

A hyphen '-' is pronounced *minus* in two cases:

1. if followed by a digit and no other digit is found in front of the hyphen, i.e. as in the pattern -X but not in X-Y or X -Z where X, Y, and Z are numbers.

2. if followed by a digit and an equals sign '=', i.e. as in the pattern X-Y=Z. Space is allowed between digits, hyphen and equals sign.

If there is no equals sign, as in X-Y or X -Z, the hyphen is pronounced *dash*.

[not SP] In certain date formats, in between days or years, the hyphen is pronounced *to*. In other cases the hyphen is never pronounced.

Expression	Reading	
-3	minus three	
44-3	forty-four dash three	
44-3=41	forty-four minus three equals forty-one	
44 - 3 = 41	forty-four minus three equals forty-one	
15-20 October	the fifteenth to twentieth of October	[not SP]
6-10 Nov	the sixth to tenth of November	[not SP]
1998-2004	nineteen ninety-eight to two thousand and four	[not SP]
02-02-2002	the second of February two thousand and two	
low-income	low income	
mother-in-law	mother in law	

4.3.2. Asterik

Asterik '*' is pronounced *multiplied by* if enclosed by digits and followed by equals sign '='. In other cases it is pronounced *asterik*.

Expression	Reading
2*3	two asterik three
2*3=6	two multiplied by three equals six
*bc	asterik b c

Chapter 5. Number Processing

Strings of digits that are sent to the text-to-speech converter are processed in several different ways, depending on the format of the string of digits and the immediately surrounding punctuation or non-numeric characters. To familiarise the user with the various types of formatted and non-formatted strings of digits that are recognised by the system, we provide below a brief description of the basic number processing along with examples. Number processing is subdivided into the following categories:

Full number pronunciation
Leading zero
Decimal numbers
Currency amounts
Ordinal numbers
Arithmetic operators
Mixed digits and letters
Time of day
Years
Dates
Phone numbers

5.1. Full number pronunciation

Full number pronunciation is given for the whole number part of the digit string.

Example

2425	full number
2,425	full number
2 425	full number
24.25	24 is a full number, 25 is the decimal part

In Indian English one lakh is used to denote one hundred thousand and one crore to denote ten million.The number grouping is based on grouping by two decimal places, rather than the three decimal places common place in most parts of the world. For example, 30 million (3 crore) rupees would be written as Rs.3,00,00,000, with commas at the thousand, lakh and crore levels instead of Rs.30,000,000.

The highest number read is 999999999999 (twelve digits). Numbers higher than this are read as separate digits.

Number	Reading
1	one
10	ten
100	hundred
1,000	one thousand
1,00,000	one lakh
1,00,00,000	one crore
1,00,00,00,000	one hundred crore
1234567890123	one two three four five six seven eight nine zero one two three

5.2. Leading zero

Numbers that begin with *0* (zero) are read as a whole number, with a zero preceding it.

Number	Reading
09253	zero nine thousand two hundred and fifty-three
020	zero twenty

5.3. Decimal numbers

Comma or full stop may be used when writing decimal numbers.

The full number part of the decimal number (the part before comma or full stop) is read according to the rules in the section *Full number pronunciation* . The decimals (the part after comma or full stop) are read as separate digits. Note: A number containing a comma followed by exactly three digits is not read as a decimal number but as a full number, following the rules in the section *Full number pronunciation* .

Number	Reading
16.234	sixteen point two three four
3.1415	three point one four one five
1251.04	one thousand two hundred and fifty-one point zero four
1,251.04	one thousand two hundred and fifty-one point zero four
2,50	two comma five zero
2.50	two point five zero
3,141	three thousand one hundred and forty-one

5.4. Currency amounts

The following principles are followed for currency amounts:

• Numbers with zero or two decimals preceded or followed by either the currency markers *£, $, ¥* or *€* or [not SP] the abbreviations *EUR, USD, DM* or *INR* are read as currency amounts.

• [not SP] Numbers with zero or two decimals followed by the words *pounds, dollars, yen* or *euros* (singular or plural) are read as currency amounts.

• Accepted decimal markers are comma ',' and full stop '.'.

• The decimal part (consisting of two digits) in currency amounts is read as *and nn pence*, *and nn cents*, or *and nn pfennigs* respectively.

• If the decimal part is *00* it will not be read.

Example	Reading	
$15.00	fifteen dollars	
15.00 USD	fifteen dollars	[not SP]
15.00£	fifteen pounds	
15.00 euros	fifteen euros	[not SP]
15.00 EUR	fifteen euros	[not SP]
€ 200.50	two hundred euros and fifty cents	

Example	Reading	
€1000	one thousand euros	
€10000	ten thousand euros	
€100 000 and €100000	one lakh euros	
€ 10,00,000 and € 1000000	ten lakh euros	[not SP]
15.45 DM	fifteen deutschmarks and forty-five pfennigs	[not SP]
Rs. 1000	one thousand rupies	[not SP]
Rs.100000 and Rs. 1 00 000	one lakh rupees	[not SP]
Rs. 10 00 000 and Rs. 1000000	ten lakhs rupees	[not SP]
Rs. 10 000 000	one crore rupees	[not SP]

There is also the possibility of writing large amounts as follows:

$ 1 million	one million dollars
¥ 1 million	one million yen

5.5. Ordinal numbers

Numbers are read as ordinals in the following cases:

• [not SP] The number is followed by a month name or one of the month name abbreviations and the number is smaller or equal to 31. The number may be preceded by a day or an abbreviation for a day.

• [not SP] The number consists of a day interval followed by a month name/abbreviation.

• The number is followed by *st, nd, rd, th, d*.

[not SP] The valid abbreviations for months are: *Jan, Feb, Mar, Apr, Jun, Jul, Aug, Sept, Oct, Nov,* and *Dec*.

[not SP] The valid abbreviations for days are: *Mon, Tue, Wed, Thu, Thurs, Fri, Sat,* and *Sun*.

The abbreviations above are only expanded to names of months and days when appearing in correct date contexts.

Expression	Reading
not SP 3 January	the third of January
not SP 3 Jan	the third of January
not SP Tuesday 3 Jan	Tuesday the third of January
not SP 15-16 January	the fifteenth to sixteenth of January
2nd May	the second of May
not SP 4th Jun 2007	the fourth of June 2007
the 21st Century	the twenty-first century
her 22nd novel	her twenty-second novel
in 3rd place	in third place
a 77th birthday party	a seventy-seventh birthday party

5.6. Arithmetic operators

Numbers together with arithmetical operators are read according to the examples below.

Expression	Reading
-12	minus twelve
14-2	fourteen dash two
14-2=12	fourteen minus two equals twelve
+24	plus twenty-four
2+3	two plus three
2+3=5	two plus three equals five
2*3	two asterisk three
2*3=6	two multiplied by three equals six
2/3	two thirds
6/2=3	six divided by two equals three
25%	twenty-five percent
3.4%	three point four percent

5.7. Mixed digits and letters

If a letter appears within a sequence of digits, the groups of digits will be read as numbers according to the rules above. The letter marks the boundary between the numbers. The letter will also be read.

Expression	Reading
77B84Z3	seventy-seven B eighty-four Z three
0092B87-B	zero zero ninety-two B eighty-seven B

5.8. Time of day

The colon is used to separate hours, minutes and seconds. [not SP]Abbreviations such as *A.M.* and *P.M.* (possible variants: *a.m., am, AM, p.m., pm, PM*) may follow or precede the time, with a space inserted between the time and the abbreviation.

[not SP] In pattern *a* below, the letter *h or H* may replace colon. Full stop is also a valid separator if one of the mentioned abbreviations is used.

[not SP] Time intervals can be denoted using a hyphen: *8-10 pm*.

Possible patterns are:

a. *hh:mm* or *h:mm*

b. *hh:mm:ss* or *h:mm:ss*

c. *[not SP] hh:mm'ss"* or *h:mm'ss"*

   Example: 12:30'45"

h = hour, *m* = minute, *s* = second.

In pattern a:

If the *mm*-part is equal to *00*, this part will not be read. Instead, *o'clock* will be added if the hours are less than 13, or *hundred hours* will be added if the hours are greater than or equal to 13.

Expression	Reading
9:00 or [not SP] 9h00	nine o'clock
9:30 pm, [not SP] 9h30 pm, or [not SP] 9.30 pm	nine thirty p m
13:00 or [not SP] 13h00	thirteen hundred hours
12:00 or [not SP] 12h00	twelve p.m.
0:00 or [not SP] 0h00	twelve a.m.

In pattern b:

An *and* will be inserted before the *ss*-part, and *seconds* will be added after it. If the *ss*-part is equal to *00*, this part will not be read.

Expresion	Reading
10:24:00	ten twenty-four
10:24:00 A.M.	ten twenty-four a m
10:24:20	ten twenty-four and twenty seconds

In pattern c:

Pattern (c) follows the rules for pattern (b).

5.9. Year

Numbers between 1100 and 1999 are always read as hundreds (year reading) with the exception of numbers containing decimals. Years (2 or 4 digits) can also be followed by *s* or *'s* to indicate decades.

Example	Reading	
1988	nineteen eighty-eight	[not SP]
1939-45	nineteen thirty-nine to forty-five	[not SP]
1088	one thousand eighty-eight	
1988.0	one thousand nine hundred and eighty-eight point zero	
1988.32	one thousand nine hundred and eighty-eight point three two	
September 1999	September nineteen ninety-nine	[not SP]
1980s	nineteen eighties	[not SP]
70's	seventies	
1980's	nineteen eighties	[not SP]

5.10. Dates

The valid formats for dates are:

1. *dd-mm-yyyy*, *dd.mm.yyyy*, and *dd/mm/yyyy*

2. *dd-mm-yy*, *dd.mm.yy*, and *dd/mm/yy*

yyyy is a four-digit number, *yy* is a two-digit number, *mm* is a month number between 1 and 12 and *dd* a day number between 1 and 31. Hyphen, full stop, and slash may be used as delimiters. In all formats, one or two digits may be used in the *mm* and *dd* part. Zeros may be used in front of numbers below 10.

Examples of valid formats and their readings:

Type 1:	Reading
10-02-2003 or 10-2-2003	the tenth of February two thousand and three
10.02.2003 or 10.2.2003	"
10/02/2003 or 10/2/2003	"

Type 2:	Reading
10-02-03 or 10-2-03	the tenth of February two thousand and three
10.02.03 or 10.2.03	"
10/02/03 or 10/2/03	"

[not SP] Ranges of days and years are also supported.

Expression	Reading
1998-1999	nineteen ninety-eight to nineteen ninety-nine
1939-45	nineteen thirty-nine to forty-five
2002/3	two thousand two to three
14-15 January	the fourteenth to fifteenth of January
October 19-20	October the nineteenth to twentieth

[not SP] Other possible formats include:

Expression	Reading
Monday, 15 January	Monday the fifteenth of January
Monday 15 January	Monday the fifteenth of January
Mon, January 15	Monday January the fifteenth
Mon January 15	Monday January the fifteenth
19 April 2007	the nineteenth of April two thousand and seven
April 19 2007	April the nineteenth two thousand and seven
May 1953	May nineteen fifty-three
3 May	third of May

5.11. Phone numbers

In this section the patterns of digits that are recognized as phone numbers are described. In the pronunciation of phone numbers, all numbers are read out digit by digit with pauses between groups of numbers. The abbreviations *tel*, *mob* and *ph* can be used in front of all the formats recognized by the system

Sequences of digits in the following formats are treated as phone numbers.

The following sequences of digits can be separated by a space or a hyphen:

Format	Example
(*area code*) xxxxxxxx	(033) 23456789
(*area code*) xxxxxx	(0821) 2345678
(*area code*) xxxxx	(03463) 234567
area code-xxxxxxxx	033-23456789

Format	Example
area code-xxxxxxx	0821-2345678
area code-xxxxxx	03463-234567
+*international code*-*area code*-xxxxxxxx	+91-33-23456789
+*international code*-*area code*-xxxxxxx	+91-821-2345678
+*international code*-*area code*-xxxxxx	+91-3463-234567
tel (*area code*) xxxxxxx	tel (0821) 2345678
mob *area code*-xxxxxxxx	mob 033-23456789
ph *international code*-*area code*-xxxxxx	ph 0091-3463-234567

The *area code* is a sequence of *0* followed by 2 to 4 digits.

Chapter 6. How to change the pronunciation

Words that are not pronounced correctly by the text-to-speech converter can be entered in the user lexicon (see *User's guide*). In this lexicon, the user enters a phonetic transcription of the word (see chapter *Indian English Phonetic Text*). Phonetic transcriptions can also be entered directly in the text, using the *PRN* tag (see *User's guide*).

Chapter 7. Indian English Phonetic Text

The Indian English text-to-speech system uses the Indian English subset of the SAMPA phonetic alphabet (*Speech Assessment Methods Phonetic Alphabet*) with some modifications. The symbols are written with a space between each phoneme.

Only the symbols listed here may be used in phonetic transcriptions. Symbols not listed here are not valid in phonetic transcriptions and will be ignored if included in the user lexicon or in a *PRN* tag.

7.1. Consonants

Table 7.1. Symbols for the Indian English consonants

Symbol	Word	Phonetic text	Comment
b	bad	b {1 d	
t	stop tomorrow	s t Q1 p t @ m Q1 r o:	
t_h	top	t_h Q1 p	
p	sport potato	s p Q:1 t p @ t_h e:1 t o:	
p_h	pad	p_h {1 d	
d	date	d e:1 t	
k	scone campaign	s k Q1 n k { m p_h e:1 n	
k_h	cone	k_h o:1 n	
g	gag	g {1 g	
m	man	m {1 n	
n	nose	n o:1 z	
r	rose	r o:1 z	
l	let	l e1 t	
L	nostril	n Q1 s t r i L	
N	ring	r I1 N	
f	fat	f {1 t	
v	vote	v o:1 t	
s	sat	s {1 t	
z	zoo	z u:1	
S	shin	S I1 n	
tS	chin	tS I1 n	
Z	measure	m e1 Z @ r	
dZ	gin	dZ I1 n	
D	this	D I1 s	
T	thin	T I1 n	
w	wait	w e:1 t	
j	yacht	j Q1 t	
h	hit	h I1 t	
hj	human	hj u:1 m @ n	

Symbol	Word	Phonetic text	Comment
b_h	bhosle	b_h _ o:1 s l e:	in some Indian proper names
d_h	dharam	d_h @1 r @ m	"
g_h	ghosh	g_h o:1 S	"
kS	paks	p @1 kS	"

7.2. Vowels

Table 7.2. Symbols for the Indian English vowels

Symbol	Word	Phonetic text	Comment
a:	father	f a:1 D @	
I	bit	b I1 t	
i:	neat	n i:1 t	
u:	zoo	z u:1	
@	hut	h @1 t	
U	put	p_h U1 t	
{	pat	p_h {1 t	
e	net	n e1 t	
e:	main	m e:1 n	
aI	high	h aI1	
OI	boy	b OI1	
o:	nose	n o:1 z	
aU	pout	p_h aU1 t	
3:	fur	f 3:1	
Q	dot	d Q1 t	
I@	near	n I@1	
e@	there	D e@1	
U@	tour	t_h U@1 r	
i	locally	l o:1 k @ l i	
u	punctual	p_h @1 N k tS u @ L	
r=	history	h I1 s t r= i	

7.3. Lexical stress

A lexical accent is used to indicate the level of prominence (or emphasis) of a syllable in a word. In Indian English, some words can be differentiated by the position of this lexical accent. The word *record* is an example of this since it can be both a noun (*a record: /r e1 k Q: d/*) or a verb (*to record: /r I k_h Q:1 d/*). Practically all words in Indian English have a lexical accent even if it does not always serve to differentiate between two different words. It is therefore important to include stress marks when writing phonetic transcriptions.

In the phonetic transcriptions, primary accent is indicated by the symbol */1/* placed directly after (no space) the accented vowel. Secondary accent is indicated by the symbol */2/*. Some examples:

devastating   /d e1 v @ s t e:2 t I N/

devastation   /d e2 v @ s t e:1 S n=/

7.4. Glottal stop

A glottal stop, represented by the phonetic symbol /?/, is a small sound which is often used to separate two words when the second word starts with a stressed vowel. This sound can be inserted in a transcription in order to improve the pronunciation.

7.5. Pause

An underscore /_/ in a phonetic transcription generates a small pause.

Chapter 8. Abbreviations

In the current version of the Indian English text-to-speech system, the abbreviations in the table below are recognized in all contexts. These abbreviations are mostly case-insensitive (except for those indicated below by "*") and require no full stop in order to be recognized as an abbreviation.

As previously mentioned, there are also abbreviations for the days of the week and the months (see chapter *Ordinal numbers*).

Table 8.1. Abbreviations

Abbreviation	Reading
kg	kilograms
°C	degrees Celsius
°F	degrees Fahrenheit
°K	degrees Kelvin
asap	A S A P
ans.	answer
Asstt.	assistant
Cantt.	cantonment
engg	engineering
b/f	before
cr	crore
cm	centimeters
corp	corporation
eg	for example
etc	et cetera
ft	feet
gov	governor
hr	hour
hrs	hours
ie	that is
jr	junior
km	kilometers
Km/h	kilometers per hour
mg	milligrams
ml	milliliters
mm	millimeters
mph	miles per hour
mr	mister
mrs	missis
ms	miss
mt	mount
prof	professor
sgt	sergeant

Abbreviation	Reading
sr	senior
tsp	teaspoon
vs	versus
gen	general
ltd	limited
dept	department
rd	road
av	avenue
ctrl	control
lb	pounds

Some abbreviations are expanded differently depending on their position in the sentence. For example, *dr* and *st* are expanded into *drive* and *street* if they appear after a capitalized noun. They are expanded into *doctor* and *saint* when they appear before a capitalized noun.

Example	Reading
Main st.	Main street
St John.	Saint John
Bayview dr.	Bayview drive
Dr. Jones.	Doctor Jones

m, g and *in.* are expanded only when appearing after a number.

Example	Reading	
25 m	twenty-five meters	
30 in.	thirty inches	(note that the dot is mandatory here)
45 g	forty-five grams	

Chapter 9. Web-addresses and email

Web-addresses and email-addresses are read as follows:

- *www* is read as three *w*'s spelled letter by letter.

- Full stops '.' are read as *dot*, hyphens '-' as *dash*, underscore '_' as *underscore*, slash '/' as *slash*.

- *us, uk, fr* and all the other abbreviations for countries are spelled out letter by letter.

- The @ is read *at*.

- Words/strings (including *org*, *com* and *edu*) are pronounced according to the normal rules of pronunciation in the system and in accordance with the lexicon.

String	Reading
www.acapela-group.com	w w w dot acapela dash group dot com
http://www.acapela-group.com	h t t p colon slash slash w w w dot acapela dash group dot com
smith@yahoo.uk	smith at yahoo dot u k
jane_smith@yahoo.uk	jane underscore smith at yahoo dot u k