



Language Manual

HQ, CO, and HD Arabic

Language Manual: HQ, CO, and HD Arabic

Published 4 February 2015

Copyright © 2008-2015 Acapela Group.

All rights reserved

This document was produced by Acapela Group. We welcome and consider all comments and suggestions. Please, use the *Contact* link at our website:

<http://www.acapela-group.com>

Table of Contents

1	GENERAL.....	1
2	LETTERS IN ORTHOGRAPHIC TEXT.....	2
3	PUNCTUATION CHARACTERS	3
4	OTHER NON ALPHANUMERIC CHARACTERS	4
5	NUMBER PROCESSING	6
6	HOW TO CHANGE THE PRONUNCIATION	13
7	ARABIC PHONETIC TEXT	14

1 General

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

This version of the document corresponds to the High Quality (HQ), Colibri (CO), and High Density (HD) Arabic voices.

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

Note: For efficiency reasons, 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 iOS
- Acapela TTS for Android



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*.

2 Letters in orthographic text

Characters from ٠-٩, A-Z and a-z may constitute a word. The Arabic diacritics are also considered as letters, like َ ِ ِْ ِْْ and ُ. Certain other characters are also considered as letters, notably those used as letters in European languages, i.e. ñ, ò, å, ç, é. These letters are not pronounced as in their native languages though. Characters outside of these ranges, i.e. numbers, punctuation characters and other non-alphanumeric characters are not considered as letters.

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*).

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, brackets and braces

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

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: Non-punctuation characters

Symbol	Reading
#	رَقْم
+	زَائِد
\	خَط مَائِل
	عَصَا
%	بِالْمِئْ
/	خَط مَائِل
=	يُسَاوِي
~	أَلِف مَدَّ
\$	دُولَار
£	لِيْفِر إِسْتِرْلِين
€	يُورُو
¥	يَان
§	عَلَامَة
*	نَجْمَة
@	أَت
&	وَ
-	شَرْطَة
_	تَسْطِير
'	فَصِيلُ فَوْقِ السَّطْرِ
÷	عَلَامَة تَنْسِيس
x	عَلَامَة دَرْب
<	أَصْغَر مِنْ
>	أَكْبَر مِنْ
,	عَلَامَة تَنْسِيس مُعَدَّل
^	مَرْفَع إِلَّا
°	سُكُونُ فَوْقِ السَّطْرِ
..	نَقْطَتَيْنِ فَوْقِ السَّطْرِ
,	عَلَامَة تَنْسِيس
`	عَلَامَة تَنْسِيس
{	فَتْحُ قَوْسٍ مُزَخْرَفٍ
}	إِغْلَاقُ قَوْسٍ مُزَخْرَفٍ
»	إِغْلَاقُ عَلَامَة تَنْسِيس
«	فَتْحُ عَلَامَة تَنْسِيس
"	عَلَامَة تَنْسِيس مُزْدَوِج

4.2 Symbols whose pronunciation varies depending on the context

4.2.1 Hyphen

A hyphen '-' is pronounced "ناقص" if followed by a digit. In certain date formats, in between days or years, the hyphen is pronounced "الى". In other cases the hyphen is never pronounced.

Expression	Reading
12-15	12 ناقص 15
12-15 Oct	12 الى 15 اكتوبر
1998-2004	1998,2004
02-02-2002	2 فبراير 2002

4.2.2 Asterisk

Asterisk '*' is pronounced مضروب في if enclosed by digits. In other cases it is pronounced نجمة.

Expression	Reading
2*3	3 مضروب في 2
*	نجمة

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 familiarize the user with the various types of formatted and non-formatted strings of digits that are recognized 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
- Year
- 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

Numbers denoting thousands, millions and billions (numbers larger than 999) may be grouped using space or comma (not full stop). In order to achieve the right pronunciation the grouping must be done correctly.

The rules for grouping of numbers are the following:

- Numbers are grouped in groups of three starting at the end.
- The first group in a number may consist of one, two, or three digits.
- If a group, other than the first, does not contain exactly three digits, the sequence of digits is not interpreted as a full number.
- The highest number read is 99999999999999 (twelve digits). Numbers higher than this are read “line of x” (خط من رقم).

Number	Reading
2580	الفان وخمسمئة وثمانون
2 580	"
2,580	"
25800	خمسة وعشرون الف وثمانمئة
25 800	"
25,800	"
1000000000	مليار

5.2 Leading zero

The 0 is read in the beginning of a number, and the next numbers are spelled

Number	Reading
02580	صفر اثنان خمسة تمانية صفر
020	صفر اثنان صفر

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 *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 *Full number pronunciation*.

Number	Reading
16.234	ستة عشر ألفا ومئتان واربعه وثلاثون
3.1415	ثلاثة فاصلة واحد اربعة واحد خمسة
2580.04	الفان وخمسمائة وثمانون فاصلة صفر اربعة
2,580.04	اثنان,خمسمائة وثمانون فاصلة صفر اربعة
2.20	اثنان فاصلة عشرون
2,20	اثنان فاصلة عشرون

5.4 Currency amounts

The following principles are followed for currency amounts:

- Numbers with zero, one or two decimal places preceded or followed by the currency markers £, \$, ¥ or € are read as monetary amounts.

- Numbers with zero or two decimal places followed by the *دولار, درهم, يورو, pounds, dollars, yen or euros* (singular or plural) are read as monetary amounts.
- Accepted decimal markers are comma and full stop.
- No spaces are allowed in the number.
- The decimal part (consisting of two digits nn) in monetary amounts is read as *بنس "nn" and سنس "nn"*.
- If the decimal part is 00 it will not be read.

Example

\$15.00.
15.00£.
15.00 euros.
€ 200.50

Reading

خمسة عشر دولار
خمسة عشر جني
خمسة عشر يورو
مئتان يورو و خمسون سنس

5.5 Ordinal numbers

Numbers are read as ordinals in the following cases:

- 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.

Examples:

3 January
3 Jan
Tuesday 3 Jan [not SP]

- The number consists of a day interval followed by a month name/abbreviation.

Example:

15-16 January

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

Examples:

1st	[not SP]
2nd	[not SP]
3rd	[not SP]
4th	[not SP]
23d	[not SP]

Valid abbreviations for months: *Jan, Feb, Mar, Apr, Jun, Jul, Aug, Sept, Oct, Nov, Dec*.

Valid abbreviations for days: *Mon, Tue, Wed, Thu, Thurs, Fri, Sat, Sun.*

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

5.6 Arithmetic operators

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

Expression	Reading
-12	ناقص اثنا عشر
+12	زائد اثنا عشر
2*3	اثنان مضروبة في ثلاثة
2/3	اثنان مقسومة في ثلاثة
25%	خمسة وعشرون بالمئة

5.7 Time of day

The colon is used to separate hours, minutes and seconds. Abbreviations such as "م", "ص", "A.M. and P.M. may follow or precede the time.

Possible patterns are:

- $hh:mm$ or $h:mm$
- $hh:mm:ss$ or $h:mm:ss$
- $hh:mm'ss''$ or $h:mm'ss''$ ex. 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, "بعد" or "صباحا", "مساء" will be added. "الزوال" will be added.

Examples:

9:00	التاسع صباحا
13:00	الواحد زوال
20:00	الثامنة مساء

In pattern (b):

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

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

5.8 Years

Numbers between 1100 and 2000 are always read as hundreds (year reading) with the exception of numbers containing decimals.

Expression	Reading
1988	الف وتسعمئة وثمانية وثمانون
1939-45	الف وتسعمئة وتسعة وثلاثون, خمسة وأربعون
September 1939	سبتمبر الف وتسعمئة وتسعة وثلاثون

5.9 Dates

The valid formats for dates are:

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

Type 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.

Example: Valid date formats and their readings

Type 1:

02-02-2003	or	02-2-2003	ثاني فبراير الفلن وثلاثة
02.02.2003	or	02.2.2003	"
02/02/2003	or	02/2/2003	"

Type 2:

02-02-03	or	02-2-03	ثاني فبراير الفلن وثلاثة
02.02.03	or	02.2.03	"
02/02/03	or	02/2/03	"

Ranges of days and years are also supported.

Examples:

14-15 January	رابع عشر الى خامس عشر يناير
---------------	-----------------------------

Other possible formats include:

- Monday, 15 January (with or without the comma)
- Mon, January 15 (with or without the comma)
- 30 April 1999
- May 1953

- 3 May

5.10 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.

5.10.1 Ordinary phone numbers

Sequences of digits in the following formats are treated as phone numbers. The following sequences of digits can be separated by a space, a period, or a hyphen:

- xxx xx xx xx
- xxx xxxx
- xx (xx) xxx xx xx
- (xx) xx xx xx xx xx
- xx (x) x xx xx xx xx
- xx (x) x xx xx xx xx
- xx x xx xx xx xx

The following sequences can only appear in these formats:

- (xx)-xxxx-xxx-xxx
- (xx).xxxx.xxx.xxx
- xx xxx xx xx
- x-xxx-xxx-xxxx

Other formats are preceded by an area code that can consist of 1-3 numbers, either surrounded by parenthesis or not. The groups of digits can be separated by a space, slash, hyphen, period or grouped together.

- area code+ xxx xxxx
- area code+ xxx xxx

5.10.2 International phone numbers

International phone numbers follow the pattern below:

International prefix + Country code + space or hyphen + Local number

International prefix: 00 or +

Country code: 1-3 digits

Local number: 6-12 digits

All formats included above can be preceded by an international prefix and a country code.

Examples

00966 (71).4521.521.843

6 *How to change the pronunciation*

The user can affect the way that the words are pronounced by using the user lexicon (see *User's guide*). It can be done in order to specify an alternative pronunciation of a certain word or to correct an erroneous pronunciation.

The pronunciation can be specified in two ways in the user lexicon. One method is to modify the spelling of the word and another is to write a phonetic transcription of the word (see chapter *Arabic phonetic text*). Phonetic transcription can also be entered directly in the text, using the PRN-tag (see *User's guide*).

7 Arabic phonetic text

The Arabic text-to-speech system uses the Arabic subset of the *SAMPA* phonetic alphabet (*Speech Assessment Methods Phonetic Alphabet*), with a few exceptions. The symbols /a./, /i./ and /u./ represent the emphatic variant of the vowels /a/, /i/ and /u/. To represent a long vowel or an accentuated consonant, one has to double the representative symbol (like /aa/, /ll/ or /rr/). The symbols are written with a space between each phoneme.

Only the symbols listed below 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: Symbols for the Arabic consonants

	stop (plosive)	fricative	nasal	flap	lateral
labial	b	f, v	m		
alveolar	t, d	s, z	n	r	l
alveolar velarized		s.			
palatal		S, Z			
velar	k, g	x, G			
glottal		h			
pharyngeal		X, H			
dental		T, D			
dental velarized	t., d.				
interdental velarized		z.			
uvular	q				

Note: *d.*, *t.*, *s.* and *z.* don't exist in the *SAMPA* notation.

Example: Arabic Consonants

phonemes	letters	examples		English translation
b	baa?	كَلْبِي	[kalbi]	my dog
t	taa?	تِلَاوَة	[tilawa]	reading book
T	thaa?	ثَعْلَب	[TaHlab]	a fox
Z	jiim	جَمَال	[Zamal]	beauty
X	Haa?	حَرْب	[Xarbun]	a war
x	khaa?	خَرَجَ	[xaraZa]	he went out
d	daal	دَخَلَ	[daxala]	he entered
D	dhaal	دَهَبَ	[Dahaba]	he left

phonemes	letters	examples		English translation
r	raa?	رَجُلٌ	[rajulun]	a man
z	zayn	يَزُورُ	[yazuru]	he visit
s	siin	سَمَكٌ	[samaka]	a fish
S	shiin	شَجَرَةٌ	[SaZara]	a tree
s.	Saad	صَبَقَ	[s.a.baqa]	he passed
d.	Daad	يُضْحِرُ	[jud.hiru]	it hurst
t.	Taa?	مُحِيطٌ	[muXit.]	sea
z.	Zaa?	مَظْلُومٌ	[maz.lum]	unfairly
H	9ayn	عِلْمٌ	[Hilmun]	knowledge
G	ghayn	غَابَةٌ	[Gabatun]	forest
f	faa?	فَازَ	[faaza]	he won
q	qaaf	قَلَمٌ	[qalamun]	a pencil
k	kaaf	كَلْبٌ	[kalbun]	a dog
l	laam	لَعِبَ	[laHaba]	he played
m	miim	مَاتَ	[mata]	he dead
n	nuun	نَامَ	[nama]	he slept
h	haa?	هَجَمَ	[haZama]	he attacked
w	waaw	ضَوْءٌ	[d.a.w?un]	a light
j	yaa?	يَلْعَبُ	[jalHabu]	he play
?	hamza	بئرٌ	[bi?run]	a well

For the geminated consonant, we double the phoneme.

Example:

b → bb

n → nn

l → l.l.

s. → s.s.

d. → d.d.

7.2 Vowels

fatha	a
kasra	i
Damma	u

Note: To have a long vowel just write *aa*, *ii* and *uu* respectively

Colored vowels

fatha

kasra

Damna

(after /d./ /t./ /z./ and /s./)

a.

i.

u.

7.3 Glottal stop

A glottal stop “همزة”, represented by the phonetic symbol /ʔ/, is a small sound which is often used to separate two vowels. This sound can be inserted in a transcription in order to improve the pronunciation.

7.4 Pause

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