BrightSpeech
Language Manual

Arabic

BrightSpeech Text-to-Speech Converter
Language Manual
Arabic
28 May 2004

This document was produced by Acapela Group. We welcome and consider all comments and suggestions. Please send them to:

Acapela Group
33, Boulevard Dolez
7000 MONS
Belgium

Tel : +32 (0)65 37 42 75
Fax : +32 (0)65 37 42 76

www.acapela-group.com

1 General

This document discusses certain aspects of text-to-speech processing for the Arabic BrightSpeech 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 voice Salma in BrightSpeech 1.2.

2 Letters in orthographic text

Characters from ي-ا, **A-Z** and **a-z** may constitute a word. The Arabic diacritics are also considered as letters, like ٰ ـَـُـِـْـٌـٍـّ and ّ. Certain other characters are also considered as letters, notably those used as letters in European languages, i.e. “ñ, õ, 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 5) and in abbreviations (see chapter 8).

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 <!> behaves 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.

Symbol	Reading
#	رَقَم
+	زَائِد
\	خَط مَائِل
|	عَصَا
%	بِالمِئَ
/	خَط مَائِل
=	يُسَاوِي
~	أَلِف مَدَّ
$	دُولَار
£	لِيفر إِستِرلِين
€	يُورُو
¥	يَان
§	عَلَامَة
*	نَجمَة
@	آت
&	وَ
-	شَرطَة
_	تَسطِير
‘	فَصِلَ فَوقَ السَطر
÷	عَلَامَةُ تَنسِيس
×	عَلَامَةُ ضَرب
<	أصغَر مِن
>	أكبَر مِن
’	عَلَامَةُ تَنسِيس مُعَدَّل
^	مَرفُع إِلَا
°	سُكُون فَوقَ السَطر
¨	نُقطَتَين فَوقَ السَطر
´	عَلَامَةُ تَنسِيس
`	عَلَامَةُ تَنسِيس
{	فَتحُ قَوسٍ مُزَخرَف
}	إِغلَاقُ قَوسٍ مُزَخرَف
»	إِغلَاقُ عَلَامَةُ تَنسِيس
«	فَتحُ عَلَامَةُ تَنسِيس
"	عَلَامَةُ تَنسِيس مُزدَوِج

Table 1 Non-punctuation characters

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 “to”. In other cases the hyphen is never pronounced.

Expression	Reading
12-15	12 ناقص 15
12-15 Oct	12 الى 15 اكتوبر

1998-2004	1998,2000
02-02-2002	2 فبراير 2000

4.2.2 Asterisk

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

Expression	**Reading**
2*3	2 مضروب في 3
*	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 99999999999 (eleven digits). Numbers higher than this are read as separate digits.

Number	**Reading**
2580	الفان وخمسمئة وثمانون
2 580	“
2,580	“
25800	خمسة وعشرون الف وثمانئة
25 800	“
25,800	“
1000000000	مليار

5.2 Leading zero

The 0 ( ) is read in the beginning of a Number

Number	**Reading**
02580	صفر الفان وخمسمئة وتمانون
020	صفر عشرون

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 5.1. 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 5.1.

Number	**Reading**
16.234	ستة عشر فاصلة مئتان واربعة وثلاثون
3.1415	ثلاثة فاصلة واحد اربعة واحد خمسة
2580.04	الفان وخمسمئة وتمانون فاصلة صفر اربعة
2,580.04	الفان وخمسمئة وتمانون 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 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 "و nn سنس" and "و nn بنس".
- If the decimal part is "00" it will not be read.

Example	**Reading**
$15.00.	خمسة عشر دولار
15.00£.	خمسة عشر جني
15.00 euros.	خمسة عشر يورو
€ 200.50	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.
- 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, 2nd, 3rd, 4th, 23d.

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

Valid abbreviations for days: 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.

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 عش
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 "ص ", "م ", "A.M." and "P.M." may follow or precede the time.

Possible patterns are:

a) **hh:mm (or h:mm)**

b) **hh:mm:ss (or h:mm:ss)**

c) **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, "صباحا", "مساءا" or "بعد الزوال" will be added

Example:	9:00	التاسع صباحا
	13:00	الواحد بعد الزوال
	20:00	الثامنة مساءا

In pattern b): An "and" will be inserted before the "ss"-part, and "ثانية" 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 Year

Numbers between 1100 and 2000 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.

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

5.9 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:	dd-mm-yyyy, dd.mm.yyyy, and dd/mm/yyyy		
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:	mm-dd-yy, mm.dd.yy, and mm/dd/yy		
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
January 19-20	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
- April 30 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:

00971-12-456-7894
00202-12 578 21 56
00966 (71).4521.521.843

6 How to change pronunciation errors

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 7). Phonetic translations can also be entered directly in the text, using a PRN-tag (see User's guide).

7 Arabic Phonetic Text

The Arabic BrightSpeech uses the Arabic subset of the SAMPA phonetic alphabet (Speech Assessment Methods Phonetic Alphabet), with a few exceptions. The symbol /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 have to double the representative symbol (like /aa/, /ll/ or /rr/). The symbols are written with a space between each phoneme.

In BrightSpeech 1.2 only SAMPA 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

7.1.1 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			
debtal velarized	t., d.				
interdental velarized		z.			
uvular	q				

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

Examples:

phonemes	letters	examples	english translation
b	baa?	كَلبي[kalbi]	my dog
t	taa?	تِلاوَة[tilawa]	reading book
T	thaa?	تَعلَب[TaHlab]	a fox
Z	jiim	جَمَال[Zamal]	beauty
X	Haa?	حَربٌ[Xarbun]	a war
x	khaa?	خَرَجَ[xaraZa]	he went out
d	daal	دَخَلَ[daxala]	he entered
D	dhaal	ذَهَبَ[Dahaba]	he left
r	raa?	رَجُلٌ[rajulun]	a man
z	zayn	يَزُورُ[yazuru]	he visit
s	siin	سَمَكَ[samaka]	a fish
S	shiin	شَجَرَة[SaZara]	a tree
s.	Saad	صَبَقَ[s.a.baqa]	he passed
d.	Daad	يُضهِرُ[jud.hiru]	it hurst
t.	Taa?	مُحِيط[muXit.]	sea
z.	Zaa?	مَظلُوم[maz.lum]	unfairly
H	9ayn	عِلمٌ[Hilmun]	khowledge
G	ghayn	غَابَة[Gabatun]	forest
f	faa?	فَازَ[faaza]	he won
q	qaaf	قَلَم[qalamun]	a pencil
k	kaaf	كَلب[kalbun]	a dog

l	laam	لَـعَبَ[laHaba]	he played
m	miim	مَـاتَ[mata]	he dead
n	nuun	نَـامَ[nama]	he slept
h	haa?	هَجَمَ[haZama]	he attacked
w	waaw	ضَوء[d.a.w?un]	a light
j	yaa?	يَـلـعَبَ[jalHabu]	he play
?	hamza	بـئـر[bi?run]	a well

For the geminated consonant, we double the phoneme.

Example:

b -> bb
n -> nn
l -> ll
s. -> s.s.
d. -> d.d.

7.2 Vowels

fatha a
kasra i
Damma u

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

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

fatha a.
kasra i.
Damma u.

7.3 Glottal stops

A glottal stop "همزة" , represented by the phonetic symbol /?/, 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.