Data Representation I

Sabah Sayed

Topics

1. Text Data
2. ASCII
3. Extended ASCII
4. UNICODE
5. Representing Sound
6. Sampling Technique
7. MIDI Technique

Text Data

- **Text (Character data)** is composed of letters, symbols, and numerals that are not used in arithmetic operations.
  - ✓‘b’ , ‘c’.
  - ✓‘+’, ‘-’ , ‘(‘, ‘)’, ‘%’ , ‘$’ .
  - ✓‘1’ , ‘2’ , ‘3’ , ‘4’ .
- Each character is assigned a unique bit pattern.
- Character data is represented using several types of codes including **EBCDIC**, **ASCII**, **Extended ASCII**, and **Unicode**.

EBCDIC

- **EBCDIC** (Extended Binary-Coded Decimal Interchange Code)
- is an 8-bit code used only by older mainframe computers.

ASCII Code

- **ASCII** (American Standard Code for Information Interchange) requires only **seven** bits for each character.

- ASCII is developed by the **A**merican **N**ational **S**tandards **I**nstitute (**ANSI**) in 1963, and finalized in 1968 as ANSI Standard X3.4.

- The purpose of ASCII was to provide a standard to code various symbols.

How many characters can ASCII represent?

- **ASCII** provides codes for 128(2^7) characters, including English letters (uppercase & lowercase), punctuation symbols, and numerals.
  - From $(0000000)_2$ $(00)_{16}$
  - To $(1111111)_2$ $(7F)_{16}$
  - For example, the **ASCII** code for an uppercase ***A*** is **1000001**.
- If 8-bits are used, the first bit is always set to 0.

ASCII Table

It only represents the English Alphabet

Symbol	ASCII	Hex	Symbol	ASCII	Hex	Symbol	ASCII	Hex
line feed	00001010	0A	>	00111110	3E	^	01011110	5E
carriage return	00001011	0B	?	00111111	3F	_	01011111	5F
space	00100000	20	@	01000000	40	`	01100000	60
!	00100001	21	A	01000001	41	a	01100001	61
"	00100010	22	B	01000010	42	b	01100010	62
#	00100011	23	C	01000011	43	c	01100011	63
$	00100100	24	D	01000100	44	d	01100100	64
%	00100101	25	E	01000101	45	e	01100101	65
&	00100110	26	F	01000110	46	f	01100110	66
'	00100111	27	G	01000111	47	g	01100111	67
(	00101000	28	H	01001000	48	h	01101000	68
)	00101001	29	I	01001001	49	i	01101001	69
*	00101010	2A	J	01001010	4A	j	01101010	6A
+	00101011	2B	K	01001011	4B	k	01101011	6B
,	00101100	2C	L	01001100	4C	l	01101100	6C
.	00101101	2D	M	01001101	4D	m	01101101	6D
/	00101110	2E	N	01001110	4E	n	01101110	6E
0	00101111	2F	O	01001111	4F	o	01101111	6F
1	00110000	30	P	01010000	50	p	01110000	70
2	00110001	31	Q	01010001	51	q	01110001	71
3	00110010	32	R	01010010	52	r	01110010	72
4	00110011	33	S	01010011	53	s	01110011	73
5	00110100	34	T	01010100	54	t	01110100	74
6	00110101	35	U	01010101	55	u	01110101	75
7	00110110	36	V	01010110	56	v	01110110	76
8	00110111	37	W	01010111	57	w	01110111	77
9	00111000	38	X	01011000	58	x	01111000	78
:	00111001	39	Y	01011001	59	y	01111001	79
;	00111010	3A	Z	01011010	5A	z	01111010	7A
<	00111011	3B	[	01011011	5B	{	01111011	7B
=	00111100	3C	\	01011100	5C	|	01111100	7C
	00111101	3D	]	01011101	5D	}	01111101	7D

Extended ASCII

- **Extended ASCII** is a superset of ASCII that uses **eight** bits to represent each character.
- For example, Extended ASCII represents the uppercase letter ***A*** as **01000001**.
- Using eight bits instead of seven bits allows Extended ASCII to provide codes for **256** characters.
- Uses the undefined space from 128-255 in ASCII, mapping it to various characters .

Extended ASCII Table

Extended ASCII character set.

“Hello.” in ASCII

01001000	01100101	01101100	01101100	01101111	00101110
H	e	l	l	o	.

- Use the ASCII table to write the ASCII code for the following:
  - Computer
  - 5=10/2

Why Numerals?

- Why does ASCII and Extended ASCII represent numerals?
  - Numerals that will not be used for calculations are better represented as text.
  - For example phone numbers are numerals but will not be used for any mathematical operations.
  - Such numerals are usually in the middle of text, such as an address (“5 Ahmed Zewail St.”).

Unicode

- **Unicode** is designed to represent the world commonly used languages.
- uses **sixteen** bits and provides codes for 65,000 characters.
- For compatibility, the first 128 Unicode are the same as the ASCII.
- It is a real bonus for representing the alphabets of multiple languages.
- Example: Unicode represents the letter د as **0000 0110 0010 1111** or $(\mathbf{062F})_{16}$ in hexadecimal.
- Unicode is a family of encoding methods (UTF-8, UTF-16, etc.)

Arabic Unicode

0600-06FF Arabic

ـ	ـ	ـ	ـ	ـ	ـ	ـ	ـ	ـ	ـ	ـ	ـ	ـ	ـ	ـ	ـ
0600	0601	0602	0603	0604	0605	0606	0607	0608	0609	060A	060B	060C	060D	060E	060F
ـ	ـ	ـ	ـ	ـ	ـ	ـ	ـ	ـ	ـ	ـ	؛	□	ـ	؟	
0610	0611	0612	0613	0614	0615	0616	0617	0618	0619	061A	061B	061C	061D	061E	061F
ي	ء	آ	أ	ؤ	إ	ئ	ا	ب	ة	ت	ث	ج	ح	خ	د
0620	0621	0622	0623	0624	0625	0626	0627	0628	0629	062A	062B	062C	062D	062E	062F
ذ	ر	ز	س	ش	ص	ض	ط	ظ	ع	غ	ـّ	ـْ	ٱ	ٹ	ٽ
0630	0631	0632	0633	0634	0635	0636	0637	0638	0639	063A	063B	063C	063D	063E	063F
ـ	ف	ق	ك	ل	م	ن	ه	و	ى	ي	ـً	ـٌ	ـٍ	ـَ	ـُ
0640	0641	0642	0643	0644	0645	0646	0647	0648	0649	064A	064B	064C	064D	064E	064F

Exercise

- Create a text file in notepad
- Write the word “hello”
- Save the file once as ANSI (ASCII) and another as Unicode.
- Using any hexa editor (https://hexed.it/) open the file and check the content of both.

What happens when I type?

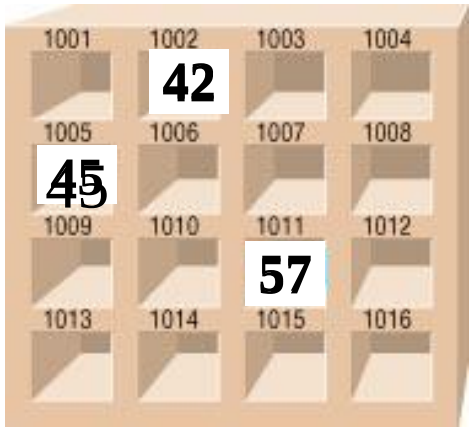
- When you type the letter A, the hardware logic built into the keyboard automatically translates that character into the ASCII code 65.
- Which is then sent to the computer.
- Similarly, when the computer sends the ASCII code 65 to output devices, the output hardware instead draw letter “A” on your screen or your computer.

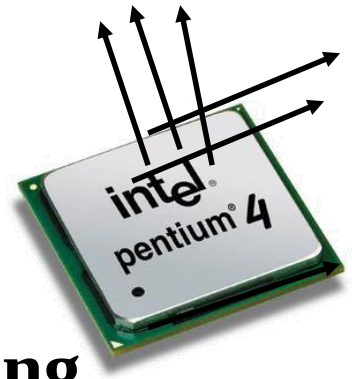
What happens when I type?

Screen

Keyboard

Memory RAM





Central Processing Unit

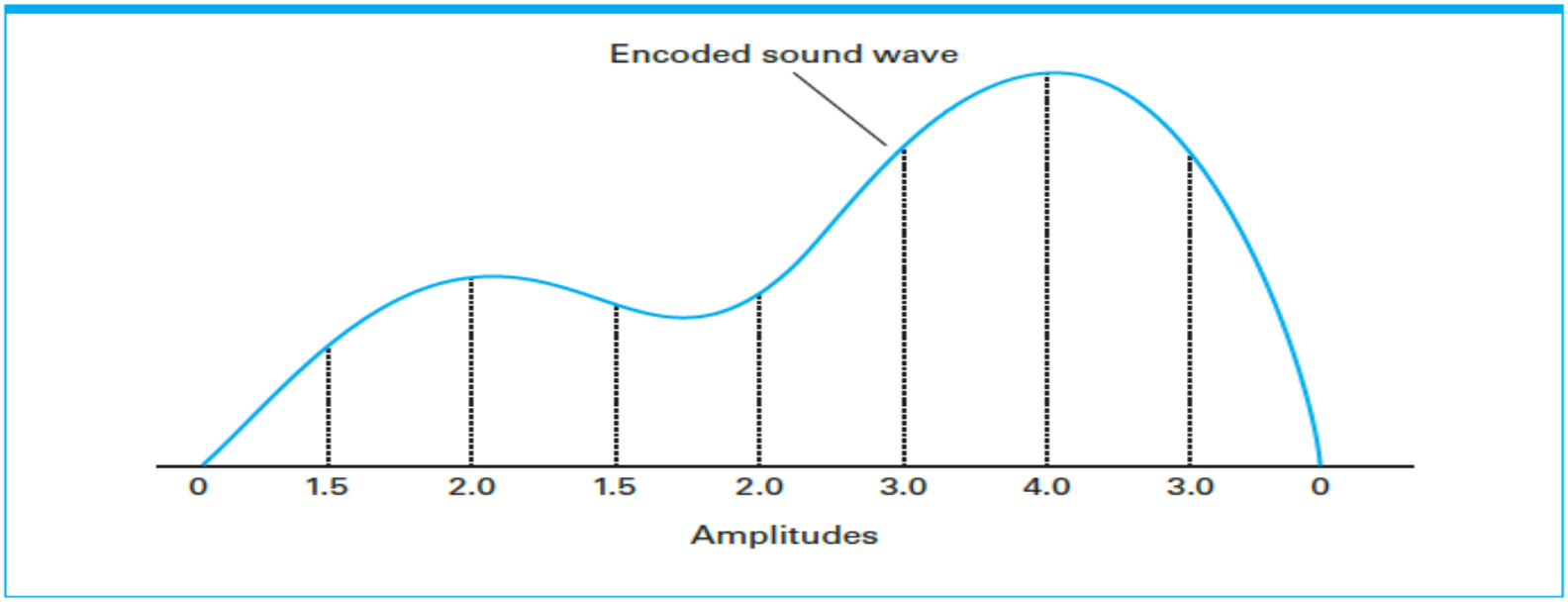
Representing Sound

- Sampling techniques
  - Used for high quality recordings
  - Records actual audio
- MIDI
  - Used in music synthesizers
  - Records “musical score”

Sampling Technique

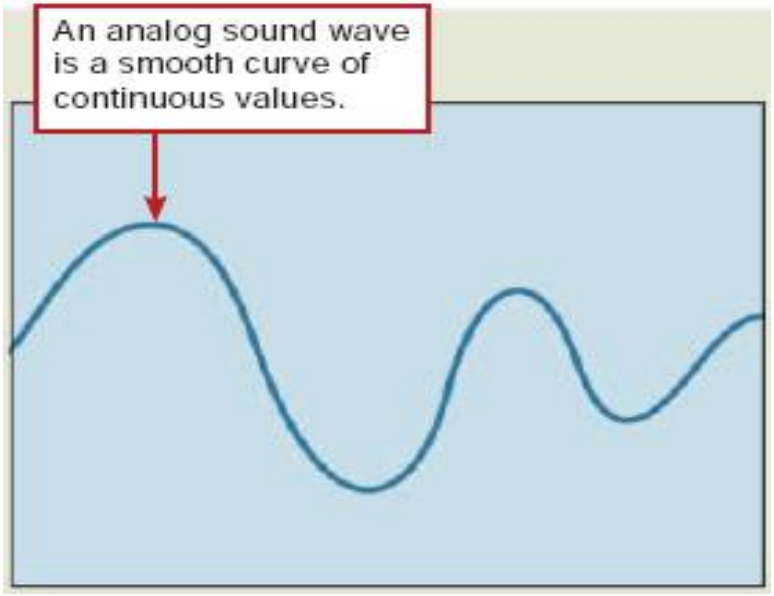
- The most generic method of encoding audio information is to sample the amplitude of the sound wave at regular intervals.

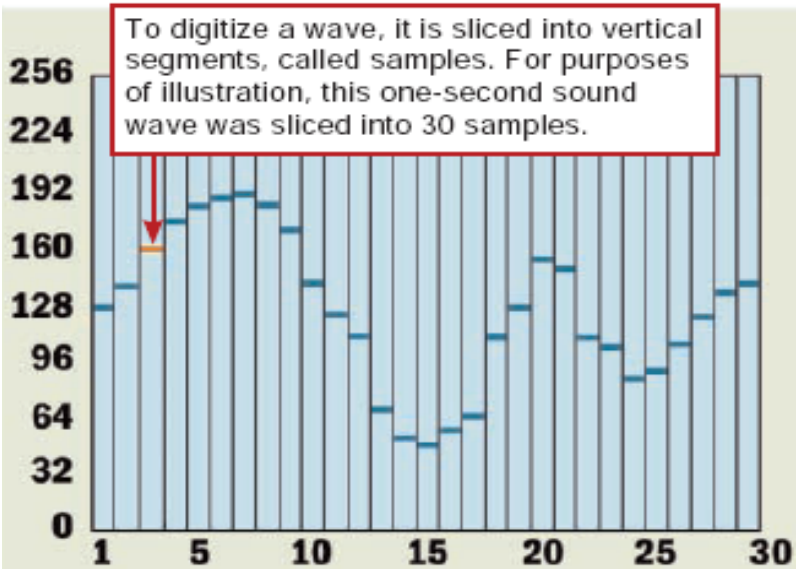
Figure 1.12 The sound wave represented by the sequence 0, 1.5, 2.0, 1.5, 2.0, 3.0, 4.0, 3.0, 0



Sampling Technique

- Digitize and then assign a binary representation





Sampling Technique

- This technique is used for voice calls with sampling rate of 8000 samples per second.
- This is not sufficient for high-fidelity music recordings.
- To obtain the quality sound reproduction obtained by today's musical CDs, a sample rate of 44,100 samples per second is used.
- The data obtained from each sample are represented in 16 bits (32 bits for stereo recordings).

MIDI Technique

- An alternative encoding system known as Musical Instrument Digital Interface is widely used in:
  - the music synthesizers
  - for video game sound
  - for sound effects accompanying websites.
- MIDI encodes what instrument is to play which note for what duration of time.
- MIDI files avoid the large storage requirements of the sampling technique.
- You can encode that a clarinet is playing the note D for two seconds in three bytes rather than more than two million bits when sampled at a rate of 44,100 samples per second.
- **Disadvantages** of MIDI files that it might sound different when performed on different synthesizers.