# Linguistics 384: Language and Computers

Topic 1: Text and Speech Encoding

Scott Martin<sup>®</sup>

Dept. of Linguistics, OSU Winter 2008

\* The course was created by Chris Brew, Markus Dickinson and Detmar Meurers.

# Writing systems used for human languages

What is writing?

"a system of more or less permanent marks used to represent an utterance in such a way that it can be recovered more or less exactly without the intervention of the utterer." (Peter T. Daniels, The World's Writing Systems)

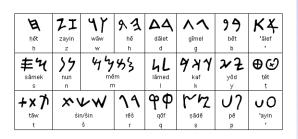
Different types of writing systems are used:

- Alphabetic
- Syllabic
- Logographic

Much of the information on writing systems and the graphics used are taken from the amazing site http://www.omniglot.com.

# Abjad example: Phoenician

An abjad used to write Phoenician, created between the 18th and 17th centuries BC: assumed to be the forerunner of the Greek and Hebrew alphabet.



(from: http://www.omniglot.com/writing/phoenician.htm)

Language and Computers – where to start?

- If we want to do anything with language, we need a way to represent language.
- We can interact with the computer in several ways:
  - write or read text
  - speak or listen to speech
- Computer has to have some way to represent
  - ▶ text speech

Alphabetic systems

Alphabets (phonemic alphabets)

- represent all sounds, i.e., consonants and vowels
- Examples: Etruscan, Latin, Korean, Cyrillic, Runic, International Phonetic Alphabet

Abjads (consonant alphabets)

- represent consonants only (sometimes plus selected) vowels; vowel diacritics generally available)
- Examples: Arabic, Aramaic, Hebrew

4/59

Language and

Computers Topic 1: Text and

Speech Encoding

Writing systems

Systems with unusual ealization

Relation to language

Encoding writter

Spoken language

Why speech is hard to represent

Relating written and

Language and

Topic 1: Text and

Speech Encoding

Writing systems

Logographic Systems with unusual

Relation to language

Encoding written

Spoken language

Why speech is hard to

Relating written and

spoken language

From Speech to Text

From Text to Speed

Writing systems

Systems with unusual realization

Relation to language

Encoding writte

Spoken language

Why speech is hard to represent

Relating written and

7/59

spoken language

From Speech to Te

From Text to Speed

language

Typing it in

Articulation

Acoustics

ASCII

Alphabet

language

Typing it in

Acoustics

ASCII

1/59

spoken language

From Speech to Text

From Text to Speech

Syllabic

Logographic

language

ASCII

Unicode

Typing it in

Articulation

Acoustics

Language and A note on the letter-sound correspondence Topic 1: Text and Speech Encoding

- Alphabets use letters to encode sounds (consonants, vowels).
- But the correspondence between spelling and pronounciation in many languages is guite complex, i.e., not a simple one-to-one correspondence.
- Example: English
  - same spelling different sounds: ought, cough, tough, through, though, hiccough
  - silent letters: knee, knight, knife, debt, psychology, mortgage
  - one letter multiple sounds: exit, use
  - multiple letters one sound: the, revolution
  - alternate spellings: jail or gaol; but chef does not have an alternative seagh (despite sure, dead, laugh)

ual	Writing systems	Alphabetic Syllabic Logographic Systems with unusual realization					
<sub>ge</sub> .tems tten	Encoding written language						
	Spoken language	language ASCII Unicode Typing it in					
age d to	Relating written and spoken language	Spoken language Transcription Why speech is hard to represent Articulation Acoustics					
en and lage <sup>24t</sup>		Relating written and spoken language From Speech to Text From Text to Speech					
2/59		3/59					
and ers	Alphabet example: Fraser						
t and coding							
ms	about 657,000 people in Myanmar, India, Thailand and in the Chinese						
	provinces of Yunnan and Sichuan.						
		Syllabic					
aual	Consonants	Logographic Systems with unusual realization					
ual ge items	P d B ſ W M M T I D S ଧ N L F J	Logographic Systems with unusual					
90	P d B ſ W M M T L D S X N L F J	Logographic Systems with unusual realization Relation to language Comparison of systems Encoding written language					
ge items	P d B ſ W M M T L D S X N L F J [p] [p'] [b] [f] [v] [m] [y] [t] [t] [d] [s] [z] [n] [i] [ts] [ts'] Z C J J X R Y K X G H A A 9 V	Logographic Systems with unusual realization Relation to language Comparison of systems Encoding written language ASCII Unicode					
ge items	P       d       B       f       W       M       M       T       L       D       S       X       N       L       F       J         [p]       [p']       [b]       [f]       [v]       [m]       [u]       [t]       [t]       [d]       [s]       [z]       [n]       [t]       [ts]       [t]	Logographic Systems with unusual realization Relation to language Comparison of systems Encoding written language ASCII					
<sup>ge</sup> items tten	$ \begin{array}{c ccccccccccccccccccccccccccccccccccc$	Lopographic Systems with unusual realization Ratation to language Comparison of systems Encoding written language ASCII Unicode Typing it in Spoken language Transcription Why speech is hard to					
<sup>ge</sup> items tten Jage	$ \begin{array}{c ccccccccccccccccccccccccccccccccccc$	Logorganic Systems with numual realization Relation to language Comparison of systems Encoding written Ianguage ASCII Unicode Typing In Spoken language Transcription					
en and lage	P       d       B       ſ       W       M       M       T       L       D       S       X       N       L       F       J         [p]       [p']       [b]       [f]       [v]       [m]       [u]       [t]       [t]       [d]       [s]       [z]       [n]       [t]       [ts]       [ts]       [t]       [ts]       [t]       [ts]       [t]       [ts]       [ts]	Lographic Systems with vusual relation to inspace Comparison of systems Encoding written Ianguage Accil Lincode Typing in Spoken Ianguage Tarsorgion Wys geneth is hard to represent Articulation Acoustics Relating written ann Spoken Ianguage					
een and	$ \begin{array}{c ccccccccccccccccccccccccccccccccccc$	Logopyte: Systems with vusual natatation Relation to Inspace Comparison of systems Encoding written Language Asci Uvicode Typing it in Spoken Language Transcription With speech is that to impresent Articulation Acoustics Relating written and					
en and tage	P       d       B       ſ       W       M       M       T       L       D       S       X       N       L       F       J         [p] [p^{h}] [b] [f] [b] [f] [v] [m] [u] [t] [t] [t] [t] [t] [t] [t] [t] [t] [t	Logopaphi Systems with vusual natatation Relation to longape Comparison of systems Encoding written Ianguage ASCII Unicode Typing kin Spoken language Taxosoption Mity speach is lard to expressent Anticulation Acoustics Relating written and spoken fair					
en and tage	$ \begin{array}{c ccccccccccccccccccccccccccccccccccc$	Logopaphi Systems with vusual natatation Relation to longape Comparison of systems Encoding written Ianguage ASCII Unicode Typing kin Spoken language Taxosoption Mity speach is lard to expressent Anticulation Acoustics Relating written and spoken fair					
en and tage	P       d       B       ſ       W       M       M       T       L       D       S       X       N       L       F       J         [p] [p^{h}] [b] [f] [b] [f] [v] [m] [u] [t] [t] [t] [t] [t] [t] [t] [t] [t] [t	Logopaphi Systems with vusual natatation Relation to longape Comparison of systems Encoding written Ianguage ASCII Unicode Typing kin Spoken language Taxosoption Mity speach is lard to expressent Anticulation Acoustics Relating written and spoken fair					

correspondences

French

Language and

Computers

Topic 1: Text and

Speech Encoding

Writing system:

Syllabic

Logographic

Systems with un realization Relation to langu

Comparison of

Encoding

language

ASCII

Linicode

Typing it in

Spoken lang

Why speech is h represent

Relating writ

spoken lang

From Speech to 1

From Text to Spe

Languag

Topic 1: Te

Speech En

Writing syste

Alphabetic Syllabic

Logographic Systems with realization

Relation to lang

Encoding v

language

Typing it in

Spoken land

Why speech is h represent

Relating writ

spoken lang

From Speech to

From Text to Spe

Language Compute

Topic 1: Text and

Speech Encoding

Writing systems

Alphabetic

Logographic

language

ASCII

Linicode

Typing it in

Articulation

Acoustics

Systems with un realization

Relation to language

Encoding writter

Spoken language

Why speech is hard to represent

Relating written and

8/59

spoken language

From Speech to Te

From Text to Speech

Articulation

Acoustics

ASCII Linicode

Corr

Comput

Articulation

Acoustics

Outline

- (1) a. Versailles  $\rightarrow$  [versai]
- b. etc, etais, etait, etaient  $\rightarrow$  [ete]

Irish

- (2) a. Baile A'tha Cliath (Dublin)  $\rightarrow$  [bl'a: kli uh]
- b. *samhradh* (summer)  $\rightarrow$  [sauruh]
- c. *scri'obhaim* (I write)  $\rightarrow$  [shgrirm]

What is the notation used within the []?

ASCII Unicode Typing it in Spoken language Why speech is hard to represent Articulation Acoustics

Language and

Computers

Topic 1: Text and

Speech Encoding

Writing system:

Topic 1: Text and

Speech Encoding

Writing systems

Alphabeti

Logographic

Systems with realization

language

Relation to language

Encoding writte

Relating written and spoken language From Speech to Tex From Text to Speed

# The International Phonetic Alphabet (IPA)

- Several special alphabets for representing sounds have been developed, the best known being the International Phonetic Alphabet (IPA).
- The phonetic symbols are unambiguous:
  - designed so that each speech sound gets its own symbol,
  - eliminating the need for
    - · multiple symbols used to represent simple sounds
    - one symbol being used for multiple sounds.
- Interactive example chart: http://web.uvic.ca/ling/ resources/ipa/charts/IPAlab/IPAlab.htm

# Syllabic alphabet example: Lao

Script developed in the 14th century to write the Lao language, based on an early version of the Thai script, which was developed from the Old Khmer script, which was itself based on Mon scripts.

### Example for vowel diacritics around the letter k:

ກ <mark>ະ</mark>	ົ້ິ	ຄຸ	ຄິ	<mark>ກ</mark> າ	ຄີ	ຄື	ຄື	<mark>ເ</mark> ກະ	ແກະ
ka	ki	ku	ku'	ka:	ki:	ku:	ku:"	ke	kae
[ka]	[ ki ]	[ku]	[ kɯ ]	[ka:]	[ ki: ]	[ku:]	[ku:]	[ke]	[kae]
ໃກະ	<mark>ເ</mark> ຄ	<mark>ແ</mark> ລ	ໃກ	<mark>ເລາະ</mark>	<mark>ເ</mark> ິ	ເຄັ້ຽ	റെല്	ຄົວ	ເກີຽ
ko	ke:	kae:	ko:	ko'	koe	kia	kia	kua	koe:y
[ ko ]	[ ke: ]	[kæ]	[ ko: ]	[kp]	[ k¥ ]	[ kiə ]	[ kia ]	[ kuə ]	[ kx:j ]
ເຄີຢ	ຄໍ	ເຄີ	ເຄືອ	ເຄົາ	<mark>ໃ</mark> ຄ	<mark>ໄ</mark> ຄ	ກໍ <mark>າ</mark>	ຄໍ	
koe:y	ko':	koe:	ku'a	kaw	kay	kay	kam	k	
[ k¥:j ]	[kɔ:]	[ k¥: ]	[kwa]	[kaw]	[ kaj ]	[ kaj ]	[kam]	[k]	

(from: http://www.omniglot.com/writing/lao.htm)

### Semantic-phonetic compounds



### An example from Ancient Egyptian

	Å.		8	500	122
msh (crocodile)	m	s	h	crocodile	m i w cat

# Syllabic systems

Language and

Computers

Topic 1: Text and

Speech Encoding

Writing systems

Systems with unusual ealization

Relation to language

Encoding written

Spoken language

Why speech is hard to represent

Relating written and

Language and

Speech Encoding

Writing systems

Logographic Systems with unusual

Relation to language

Encoding written

Spoken language

Why speech is hard to

Relating written and

Language and Computers

Topic 1: Text and

Speech Encoding

Writing systems

tems with unusual

Relation to language

ncoding writte

Spoken language

Why speech is hard to represent

Relating written and

16/59

spoken language

From Text to Speed

language

Typing it in

Acoustics

= cat (miw)

ASCII

13/59

spoken language

From Speech to Tex

From Text to Speed

language

Typing it in

Acoustics

ASCII

Co Topic 1: Text and

spoken language

From Speech to Text

From Text to Speech

Comparison of sys

language

ASCII

Unicode

Typing it in

Articulation

Acoustics

Alphabetic

Logographic

Syllabic alphabets (Alphasyllabaries)

- writing systems with symbols that represent a consonant with a vowel, but the vowel can be changed by adding a **diacritic** (= a symbol added to the letter).
- Examples: Balinese, Javanese, Tibetan, Tamil, Thai, Tagalog

(cf. also: http://www.omniglot.com/writing/syllabic.htm)

### Syllabaries

- writing systems with separate symbols for each syllable of a language
- Examples: Cherokee. Ethiopic, Cypriot, Ojibwe, Hiragana (Japanese)

(cf. also: http://www.omniglot.com/writing/syllabaries.htm#syll)

# Logographic writing systems

- Logographs (also called Logograms):
  - Pictographs (Pictograms): originally pictures of things, now stylized and simplified.

Example: development of Chinese character horse:



- Ideographs (Ideograms): representations of abstract ideas
- Compounds: combinations of two or more logographs
- Semantic-phonetic compounds: symbols with a meaning element (hints at meaning) and a phonetic element (hints at pronunciation).
- Examples: Chinese (Zhōngwén), Japanese (Nihongo), Mayan, Vietnamese, Ancient Egyptian

Two writing systems with unusual realization

## Tactile

- Braille is a writing system that makes it possible to read and write through touch; primarily used by the (partially) blind.
- It uses patterns of raised dots arranged in cells of up to six dots in a 3 x 2 configuration.
- Each pattern represents a character, but some frequent words and letter combinations have their own pattern.

### Chromatographic

The Benin and Edo people in southern Nigeria have developed a system of writing based on different color combinations and symbols.

(cf. http://www.library.cornell.edu/africana/Writing\_Systems/Chroma.html

#### Language and Syllabary example: Cypriote Computers Topic 1: Text and

(from: http://www.omnialot.com/writina/cvpriot.htm)

Speech E

Writing syst

Syllabic

Systems with u realization Relation to lan

Comparison of

Encoding

language

ASCII

Unicode

Typing it in

Spoken lar

Why speech is represent

Articulation Acoustics

Relating w

spoken lan

From Speech to From Text to Spe

Language a

Topic 1: Text a

Speech Encod

Writing system

Relation to langua

Encoding writte

language

Typing it in

Articulation

Acoustics

Spoken langua

Why speech is hard represent

Relating writter

spoken languag

From Speech to Te

From Text to Speech

14/59

Topic 1: Text and

Speech Encoding

Writing systems

llabic

language

Typing it in

Articulation

Acoustics

ASCII

terns with ization

Relation to language

Encoding writter

Spoken language

Why speech is hard to represent

Relating written and

spoken language

From Text to Speed

ASCII

Alphabeti

Computer

Logographic

Language and Computers Topic 1: Text and eech Encodin

Encoding	The	Cypri	iot syll	abary	or Cy	pro-M	inoan	writing	g is the	ought	to hav	e		Speech Encoding
stems	deve	eloped	d from	the Li	near /	A, or p	ossibl	y the L	inear	B scri	pt of (	Crete,		Writing systems
														Alphabetic
	thou	igh its	s exact	i origir	is are	not kr	iown.	lt was	used	from a	about	800 to	200	Syllabic
	<b>D</b> O													Logographic Systems with unusual
unusual	BC.													realization
guage														Relation to language
f systems	Ж	F	١X	$\uparrow$	+	$\sim$	X	T	0	V	Х	)(	0	Comparison of systems
written			<i>/</i> \		•		· · ·		<u>_v</u> _	۷		/\		Encoding written
	а	ta	ga	ka	pa	la.	ma	na	ra	sa	vа	ха	ya	language
	Ж			×	4	8	Ж	151	$\sim$	ш	т	(]		ASCII
		$\mathbf{Y}$		×	/	0	∕∧	171	11	1	1	(		Unicode
	e	te		ke	pe	le	me	ne	re	se	ve	хе		Typing it in
nguage	$\mathbf{V}$	$\uparrow$		$\overline{\mathbf{n}}$	$\leq$	/	$\sim$	Ł	Ŷ	$\wedge$	N/Z			Spoken language
				Τ.	~	=			$\checkmark$	<u> </u>	/ \			Transcription
s hard to	i	ti		ki	pi	li	mi	ni	ri	si	vi			Why speech is hard to
	$\sim$	T		П	۶	1	-	7.	0	84	$\wedge$	U		represent
	Ŷ			11	7	T	Φ	71	X	$\leq$	J	- 55	$\sim \sim$	Articulation
	0	to		ko	ро	k	mo	no	ro	so	YO	zo	yo	Acoustics
ritten and	$\Upsilon$	Τī		×	$\mathbf{v}$	M	$\bowtie$	Х	).(	Ж				Relating written and spoken language
to Text	ů	tu		ku	pu	lu l	mu	nu	'n	su	yu			From Speech to Text
ipeech					1- M					24				From Text to Speech

1/59		_
und s	Logograph writing system example: Chinese	
and ding	Pictographs	
IS	女子口日月山川豕目心雨田木龜 woman child mouth sun moon mountain (iver pig eve heart rain field tree turtle	V S
al ms	Ideographs	
en	一二三上下中力凸凹 one two three above below middle stendth convex concave	E la u
ige	one two three above below middle stength convex concave (pleugh)	S 1
	Compounds of Pictographs/Ideographs	,
n and	好安明家思牢雷男	F
ge	good peaceful bright home/family thought prison thunder man/male	S
	(woman (woman under (sun + moon) (pig under (heart + field) (cow under (rain cloud (field + strength)	

a roof

a roof)

over a field

#### Language and Computers Braille alphabet

+ child)

a roof)

(from: http://www.omnialot.com/writing/chinese\_types.htm

•	:	••	.:	••	:.	::	:.		.:	:	:	:
Α	В	С	D	E	F	G	Н	1	J	K	L	M
1	2	3	4	5	6	7	8	9	0			
a	but	can	do	every	from	go	have		just	knowledge	like	mor
::	÷	÷	÷	÷	:	::	:	1.	•		::	::
Ν	0	Ρ	Q	R	S	Т	U	V	W	×	Y	Ζ
not		people	quite	rather	S0	that	us	very	will	t.	you	86
::	ii	::	::	::	•.	:.	::	:	:	::	÷	:
Ç	É	À	È	Ù	Â	Ê	Î	Ô	Û	Ë	Ĩ	Ü
and	for	of	the	with	child		shall	this	which			out
				<u> </u>	ch	gh	sh	th	wh	ed	er	ou
÷	•	:	••	::	٠.	:.	::	:.		.:		
ÓŒ		)	1			1	()	?"	*		Ì	Ò
											fraction line	
011		bb	cc	did	en		gg; were		in		st	ing
.:	.:	•	••	•	:	:	:	•				
	ĂÆ	'	-	numerical	literal	italic sign	letter	capital				
numeral				index	index	decimal	sign	sign				
sign	ar			accent		sign			J			

Articulation Acoustics Relating written an spoken language From Text to Speed

Language and Topic 1: Text and Speech Encoding Writing system:

lelation to language Encoding written anguage ASCII Unicode Typing it in Spoken language

Why speech is hard to Acoustics Relating written and spoken language n Speech to Text

From Text to Speed

Language and

Topic 1: Text and

Speech Encoding

Writing systems

Relation to language

Encoding writter

Spoken language

Why speech is hard to represent

Alphabeti

Syllabic

Compari

language

ASCII

Unicode

Typing it in

# Chromatographic system



# Japanese example

The example uses kanji (red), hiragana (black), and katakana (blue):

カプヤルホテル 各室がカプセル形の簡易ホテル。終電に乗り遅れたサラリーマンなどが高いタクシ ー代を払って帰宅するより安く済むことから、手軽に利用している。

#### Translation:

#### Capsule Hotel

A simple hotel where each room is capsule-shaped. When businessmen miss the last train home, they can stay overnight very cheaply instead of paying a lot of money to go home by taxi

(from: http://www.omniglot.com/writing/iapanese.htm#origin)

# Comparison of writing systems

What are the pros and cons of each type of system?

- accuracy: Can every word be written down accurately?
- Iearnability: How long does it take to learn the system?
- cognitive ability: Are some systems unnatural? (e.g. Does dyslexia show that alphabets are unnatural?)
- Ianguage-particular differences: English has thousands of possible syllables; Japanese has very few in comparison
- connection to history/culture: Will changing a writing system have social consequences?

Relating writing systems to languages

- There is not a simple correspondence between a writing system and a language.
- For example, English uses the Roman alphabet, but Arabic numerals (e.g., 3 and 4 instead of III and IV).
- We'll look at three other examples:
  - Japanese
  - Korean
  - Azeri

#### Language and Korean

19/59

Language and

Computers Topic 1: Text and

Speech Encoding

Writing systems

ation to language

Encoding written

Spoken language

Why speech is hard to represent

Relating written and

Topic 1: Text and

Speech Encoding

Writing systems

Logographic Systems with unusual

Relation to language

Encoding written

Spoken language

Why speech is hard to represent

Relating written and

22/59

Language and Computers

Topic 1: Text and

Speech Encoding

Writing systems

Systems with unusual realization

Why speech is hard to represent

Relating written and

25/59

spoken language

From Speech to Ter

From Text to Speech

Acoustics

Syllabic

spoken language

From Speech to Text

From Text to Speech

language

ASCII

Unicode

Typing it in

Articulation

Acoustics

spoken language

From Speech to Text

From Text to Speech

language

ASCII

Unicode

Typing it in

Articulation

Acoustics

Syllabic

"Korean writing is an alphabet, a syllabary and logographs all at once." (http://home.vicnet.net.au/~ozideas/writkor.htm)

- ► The hangul system was developed in 1444 during King Sejong's reign.
  - There are 24 letters: 14 consonants and 10 vowels
  - But the letters are grouped into syllables, i.e. the letters in a syllable are not written separately as in the English system, but together form a single character. E.g., "Hangeul" (from: http://www.omniglot.com/writing/korean.htm):

한 (han) ㅎ(h) + ㅏ(a) + ㄴ(n) 글 (geul) ㄱ(g) + ㅡ(eu) + ㄹ(l)

► In South Korea, hanja (logographic Chinese characters) are also used.

## Encoding written language

- Information on a computer is stored in bits.
- A bit is either on (= 1, yes) or off (= 0, no).
- A list of 8 bits makes up a byte, e.g., 01001010
- Just like with the base 10 numbers we're used to, the order of the bits in a byte matters:
  - Big Endian: most important bit is leftmost (the standard way of doing things)
    - The positions in a byte thus encode: 128 64 32 16 8 4 2
    - "There are 10 kinds of people in the world; those who know binary and those who don't" (from: http://www.wlug.org.nz/LittleEndian)
  - Little Endian: most important bit is rightmost (only used on Intel machines)
    - The positions in a byte thus encode: 1 2 4 8 16 32 64 128

#### Language and Japanese Computers Topic 1: Text and Speech Encoding

Writing systems

Systems with unusual realization

Relation to language

Encoding written

Spoken language

Why speech is hard to represent

Relating written and

Language and

Topic 1: Text and

Speech Encoding

Writing system:

Logographic Systems with unusual

Relation to language

Encoding written

Spoken language

Why speech is hard to

Relating written and

23/59

Language and Computers

Topic 1: Text and

Speech Encoding

Writing systems

Systems with unusual realization

Relation to language

codina written

Spoken language

Why speech is hard to represent

Relating written and

26/59

spoken language

From Speech to Te

From Text to Speech

Comparison of sys

Syllabic

ASCII

Typing it in

Articulation

Acoustics

Logographic

spoken language

From Speech to Tex

From Text to Speech

language

Typing it in

Articulation

Acoustics

Alphabe

Azeri

spoken language

From Speech to Text From Text to Speech

Logographic

language

ASCII

Unicode

Typing it in

Articulation

Acoustics

Japanese: logographic system kanji, syllabary katakana, syllabary hiragana

- kanji: 5,000-10,000 borrowed Chinese characters
- katakana
  - used mainly for non-Chinese loan words, onomatopoeic words, foreign names, and for emphasis

#### hiragana

- originally used only by women (10th century), but codified in 1946 with 48 syllables
- used mainly for word endings, kids' books, and for words with obscure kanji symbols
- romaji: Roman characters

Why speech is hard to represent Articulation Acoustics Relating written and spoken language

Language and Topic 1: Text and Speech Encoding

A Turkish language with speakers in Azerbaijan, northwest Iran, and (former Soviet) Georgia

- ► 7th century until 1920s: Arabic scripts. Three different Arabic scripts used
- 1929: Latin alphabet enforced by Soviets to reduce Islamic influence.
- 1939: Cyrillic alphabet enforced by Stalin
- 1991: Back to Latin alphabet, but slightly different than before.

→ Latin typewriters and computer fonts were in great demand in 1991

Converting decimal numbers to binary - Tabular Method

Using the first 4 bits, we want to know how to write 10 in bit (or binary) notation.

8	4	2	1
?	?	?	?
8 < 10	?	?	?
1	8 + 4 = 12 > 10	?	?
1	0	8 + 2 = 10 = 10	?
1	0	1	0

Relating written and spoken language From Speech to Tex From Text to Speed

#### Language and Computers Topic 1: Text and Speech Encoding Writing system:

Syllabic

Logographic

#### Systems with unusual Relation to language Encoding written language Unicode Typing it is

Spoken language

From Speech to Text

From Text to Speech

Writing system: Alphabeti Logographic Relation to language

Encoding written language ASCII Unicode

Spoken language Why speech is hard to Articulation

Typing it in

Acoustics Relating written and spoken language From Speech to Text From Text to Speed

24/59

Language and Topic 1: Text and

Writing systems Alphabetic Syllabic Logographie

Speech Encoding

Systems with unusual realization Relation to language Comparison of system Encoding written

language ASCII Unicode Typing it in Spoken language

Why speech is hard to represent Articulation Acoustics

27/59





# Converting decimal numbers to binary - Division Method

Decimal	Remainder?	Binary
10/2 = 5	no	0
5/2 = 2	yes	10
2/2 = 1	no	<mark>0</mark> 10
1/2 = 0	yes	<mark>1</mark> 010

# The ASCII chart

Codes 1-31 are used for control characters (backspace, line feed, tab, ...).

32	1	48	0	65	A	82	R	97	а	114	r
33	1	49	1	66	в	83	S	98	ь	115	s
34		50	2	67	С	84	т	99	с	116	t
35	#	51	3	68	D	85	U	100	d	117	u
36	\$	52	4	69	E	86	v	101	с	118	v
37	%	53	5	70	F	87	W	102	f	119	w
38	&	54	6	71	G	88	х	103	8	120	x
39	· ·	55	7	72	н	89	Υ	104	h	121	у
40	(	56	8	73	Ι	90	Z	105	i i	122	z
41	)	57	9	74	1	91		106	j	123	1
42	*	58	:	75	K	92	Ň	107	k	124	-
43	+	59		76	L	93	1	108	1	125	}
44	÷.	60	<	77	М	94	-	109	m	126	~
45		61	=	78	Ν	95	-	110	n	127	DEL
46		62	>	79	0	96	•	111	0		
47	/	63	?	80	Р			112	р		
	-	64	@	81	Q			113	q		

# Different coding systems

But wait, didn't we want to be able to encode all languages? There are ways ...

- Extend the ASCII system with various other systems, for example:
  - ISO 8859-1: includes extra letters needed for French, German, Spanish, etc.
  - ► ISO 8859-7: Greek alphabet
  - ► ISO 8859-8: Hebrew alphabet
  - ► JIS X 0208: Japanese characters
- Have one system for everything  $\rightarrow$  **Unicode**

1	Language and Computers Topic 1: Text and Speech Encoding	Using bytes to store characters	Language and Computers Topic 1: Text and Speech Encoding	An encoding standard: ASCII
	Writing systems Abshate: Lapopapito Systems with nusael mattation to impage Comparison of systems Encoding writing Impage and a system Encoding writing ASCH Unicode Encoding writing ASCH Unicode Protection Wy specific is that to specific Ascutation Accu	<ul> <li>With 8 bits (a single byte), you can represent 256 different characters. Why would we want so many?</li> <li>If you look at a keyboard, you will find lots of non-English characters.</li> <li>With 256 possible characters, we can store every single letter used in English, plus all the things like commas, periods, space bar, percent sign (%), back space, and so on.</li> </ul>	Writing systems Aphabetic Systems with unsual realization Relation to language Comparison of systems Encoding writinen Encoding writinen Ascut Ukicods Typing in Spoken language Anstuation Assutics Relating writinen and Rechaing writinen and Prom Speech Towa Prom Speech Towa	<ul> <li>ASCII = the American Standard Code for Information Interchange</li> <li>7-bit code for storing English text</li> <li>7 bits = 128 possible characters.</li> <li>The numeric order reflects alphabetic ordering.</li> </ul>
	28/59 Language and Computers Topic 1: Text and Speech Encoding	E-mail issues	29/59 Language and Computers Topic 1: Text and Speech Encoding	Multipurpose Internet Mail Extensions (MIME)
	Writing systems Abhabaic Systems with nucual Loopopatric Loopopatric Relation tangange Comparison of systems Encoding written Ianguage Xetai Tanguage Xetai Tanguage Spoken Language Taneorstinn Wy seen is hard to Taneorstinn Wy seen is hard to Taneorstinn Wy seen is hard to Taneorstinn Rocastic Relating written and spoken Language Fron Speach to Tel Fron Speach to Tel	<ul> <li>Have you ever had something like the following at the top of an e-mail sent to you? [The following text is in the ''ISO-8859-1'' character set.] [Your display is set for the ''US-ASCII'' character set. ]</li> <li>[Some characters may be displayed incorrectly. ]</li> <li>Mail sent on the internet used to only be able to transfer the 7-bit ASCII messages. But now we can detect the incoming character set and adjust the input.</li> <li>Note that this is an example of meta-information = information which is printed as part of the regular message, but tells us something about that message.</li> </ul>	Writing systems Aphabetic System Lopographic Systems with unausal realization comparison of systems Encoding written Ianguage Monode Typing 1 Spoken Ianguage Taracrejsion Way speech is hard to represent Actuation Accustics For Speech Tart Accustics	<ul> <li>MIME provides meta-information on the text, which tells us:</li> <li>which version of MIME is being used</li> <li>what the charcter set is</li> <li>if that character set was altered, how it was altered</li> <li>Mime-Version: 1.0 Content-Type: text/plain; charset=US-ASCII</li> <li>Content-Transfer-Encoding: 7bit</li> </ul>
	Aitos Language and Concutes Topic 1: Text and Speech Encodings Vitting systems Lagopaini System Lagopaini System Concutes outputs System Concutes outputs Concutes outputs Concute	<ul> <li>Unicode</li> <li>Problems with having multiple encoding systems:</li> <li>Conflicts: two encodings can use the same number for two different characters and use different numbers for the same character.</li> <li>Hassle: have to install many, many systems if you want to be able to deal with various languages</li> <li>Unicode tries to fix that by having a single representation for every possible character.</li> <li>"Unicode provides a unique number for every character, no matter what the platform, no matter what the program, no matter what the language." (www.unicode.org)</li> </ul>	23/53	<ul> <li>How big is Unicode?</li> <li>Version 3.2 has codes for 95,221 characters from alphabets, syllabaries and logographic systems.</li> <li>Uses 32 bits – meaning we can store 2<sup>32</sup> = 4,294,967,296 characters.</li> <li>4 billion possibilities for each character? That takes a lot of space on the computer!</li> </ul>

### Language and Computers Topic 1: Text and Speech Encoding

Writing systems Alphabeti Syllabic Logographic Systems with unusual realization Relation to language Comparison of system Encoding written



Why speech is hard to represent Articulation Acoustics

Relating written and spoken language From Speech to Text From Text to Speech

# 30/59

Language and Computers Topic 1: Text and Speech Encoding

Writing systems Alphabetic Syllabic Logographic Systems with unusual realization

Relation to language Corr

Encoding written language ASCII

Why speech is hard to represent Articulation Acoustics Relating written and

spoken language From Speech to Text From Text to Speech

Unicode Typing it in Spoken language

- ing used
- red, how it was altered

33/59

### Language and Computers Topic 1: Text and Speech Encoding Writing systems Alphabeti Syllabic

Logographi Systems with u realization Relation to language Comparison of sy Encoding written language ASCII Unicode

Typing it in Spoken language Transcription Why speech is hard to represent Articulation Acoustics

Relating written and spoken language From Speech to Text From Text to Speech

# Compact encoding of Unicode characters

- Unicode has three versions
  - UTF-32 (32 bits): direct representation
  - ▶ UTF-16 (16 bits): 2<sup>16</sup> = 65536
  - ► UTF-8 (8 bits): 2<sup>8</sup> = 256
- How is it possible to encode 2<sup>32</sup> possibilities in 8 bits (UTF-8)?
  - Several bytes are used to represent one character.
  - Use the highest bit as flag:
    - highest bit 0: single character
    - highest bit 1: part of a multi byte character
  - Nice consequence: ASCII text is in a valid UTF-8 encoding.

# The need for speech

- What if we want to work with an unwritten language?
- What if we want to examine the way someone talks and don't have time to write it down?

Many applications for encoding speech:

- Building spoken dialogue systems, i.e. speak with a computer (and have it speak back).
- Helping people sound like native speakers of a foreign language.
- Helping speech pathologists diagnose problems

Articulatory properties: How it's produced

We could talk about how sounds are produced in the vocal tract, i.e. articulatory phonetics

- place of articulation (where): [t] vs. [k]
- manner of articulation (how): [t] vs. [s]
- voicing (vocal cord vibration): [t] vs. [d]

But unless the computer is modeling a vocal tract, we need to know acoustic properties of speech which we can quantify.

How do we type everything in? Topic 1: Text and

- Use a keyboard tailored to your specific language e.g. Highly noticeable how much slower your English typing is when using a Danish-designed keyboard.
- Use a processor that allows you to switch between different character systems. e.g. Type in Cyrillic characters on your English
- keyboard. Use combinations of characters.
- An e followed by an 'might result in an é Pick and choose from a table of characters.

So, now we can encode every language, as long as it's written.

#### 37/5 Language and

Language and

Computers

Speech Encoding

Writing systems

Systems with unusual ealization

Relation to language

Encoding written

Spoken language

Why speech is hard to represent

Relating written and

Topic 1: Text and

Speech Encoding

Writing systems

Logographic Systems with unusual

Relation to language

Encoding written

Spoken language

Why speech is hard to

Relating written and

spoken language

From Speech to Text

From Text to Speed

language

ASCII

Unicode

Typing it in

Articulation

Acoustics

spoken language

From Speech to Text From Text to Speech

Comparison of sys

language

Unicode

Typing it i

Articulation

Acoustics

Syllabic

Logographic

What does speech look like?

We can transcribe (write down) the speech into a phonetic alphabet.

- It is very expensive and time-consuming to have humans do all the transcription.
- To automatically transcribe, we need to know how to relate the audio file to the individual sounds that we hear.
- $\Rightarrow$  We need to know:

hear

- some properties of speech
- how to measure these speech properties how these measurements correspond to sounds we

40/59

Speech Encoding

Writing systems

Relation to language

Encoding writte

Spoken language

Why speech is hard to represent

Relating written and

43/59

spoken language

From Text to Speech

language

ASCII

Unicode

Typing it in

Transcription

Articulation

#### Language and Acoustic properties: What it sounds like Topic 1: Text and

**Sound waves** = "small variations in air pressure that occur very rapidly one after another" (Ladefoged, A Course in Phonetics)

 $\Rightarrow$  Akin to ripples in a pond

- speech flow = rate of speaking, number and length of pauses (seconds)
- Ioudness (amplitude) = amount of energy (decibels)
- frequencies = how fast the sound waves are repeating (cycles per second, i.e. Hertz)
  - pitch = how high or low a sound is
  - In speech, there is a fundamental frequency, or pitch. along with higher-frequency overtones.
- intonation = rise and fall in pitch

#### Language and Unwritten languages

Computers Topic 1: Text and Speech Encoding

Writing system: Syllabic Logographic Systems with unusual realization Relation to language

Comparison of system Encoding written language ASCII

Spoken language Why speech is hard to represent

Many languages have never been written down. Of the 6700 spoken, 3000 have never been written down. Salar, a Turkic language in China

- Gugu Badhun, a language in Australia.
- Southeastern Pomo, a language in California

Articulation Relating written and spoken language From Speech to Tex From Text to Speecl

Topic 1: Text and

Speech Encoding

Writing systems

Logographic Systems with unusual

Relation to language

Encoding written

Spoken language

41/59

Language and Computers

Topic 1: Text and

Speech Encoding

Writing systems

Systems with unusua realization

Relation to language

Encoding written

Spoken language

Why speech is hard to represent

Relating written and

spoken language

From Text to Speech

Alphabet

Syllabic

Comparie

language

ASCII

Unicode

Typing it in

Articulation

Logographic

language

ASCII

Linicode

Typing it in

Alphabe

#### 38/59 Language and

What makes representing speech hard?

### Difficulties:

- People have different dialects and different size vocal tracts and thus say things differently
- Sounds run together, and it's hard to tell where one sound ends and another begins.
- What we think of as one sound is not always (usually) said the same: coarticulation = sounds affecting the way neighboring sounds are said

e.g. k is said differently depending on if it is followed by ee or by oo.

- What we think of as two sounds are not always all that different.
- e.g. The s see is very acoustically similar to the sh in shoe

(Check out the Speech Analysis Tutorial, of the Deptartment of Linguistics at Lund University, Sweden at

http://www.ling.lu.se/research/speechtutorial/tutorial.html, from which the illustrations on this and the following

0 n ə t 1

Oscillogram (Waveform)

Writing systems Alphabet

Systems with realization Relation to language Encoding writte language ASCII Unicode

n

Spoken language Why speech is hard to Articulatio

Relating written and spoken language From Speech to Text From Text to Speed

slides are taken.)

45/59

Speech Encoding Writing system: Alphabe Logographic Systems with unusu

Language and

Topic 1: Text and

39/5

Language and

Computers

Topic 1: Text and

Speech Encoding

Writing system:

Systems with unusual realization Relation to language

Comparison of syste

Encoding written

Spoken language

Why speech is hard to represent

spoken language

From Speech to Text

From Text to Speech

Syllabic

Logographic

language

ASCII

Unicode

Typing it in

Articulation

Acoustics Relating written and

Relation to language

Encoding written language ASCII Unicode

Typing it in Spoken language

Why speech is hard to Acoustics

Relating written and spoken language From Speech to Text From Text to Speed

ſ ə

Language and Topic 1: Text and Speech Encoding

42/59

Syllabic

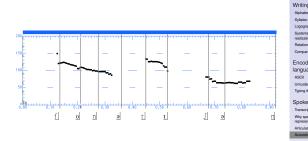
Logographie

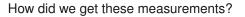
Typing it in

Why speech is hard to Articulation Acoustics

Relating written and spoken language From Speech to Tex From Text to Speed

# Fundamental frequency (F0, pitch)





**sampling rate** = how many times in a given second we extract a moment of sound; measured in samples per second

Sound is continuous, but we have to store data in a discrete manner.



· We store data at each discrete point, in order to capture the general pattern of the sound

Automatic Speech Recognition (ASR)

Automatic speech recognition = process by which the computer maps a speech signal to text. Uses/Applications:

- Dictation
- Telephone conversations
- ▶ People with disabilities e.g. a person hard of hearing could use an ASR system to get the text

#### Language and Spectrograms

Language and Computers Topic 1: Text and Speech Encoding	Spectrograms	Language and Computers Topic 1: Text and Speech Encoding	How measurements correspond to sounds we hear
Writing systems Aphaton System (System) (Comparation Appropriate Appropriate (Comparation of systems Encoding written Ianguage ASCI Ianguage (Comparison of systems (Comparison of systems) (Comparation of systems) (Compara	Spectrogram = a graph to represent (the frequencies of) spech over time.	Aphabatics Syllabic Lapopraphics Syllabics Lapopraphics Aphabatics Comparison of systems Comparison of systems Comparison of systems Comparison of systems Comparison of Systems Comparison Systems Comparison Systems Comparison Systems Comparison Comparis	<ul> <li>How dark is the picture? → How loud is the sound? We can measure this in decibels.</li> <li>Where are the lines the darkest? → Which frequencies are the loudest and most important? We can measure this in terms of Hertz, and it tells us what the vowels are.</li> <li>How do these dark lines change? → How are the frequencies changing over time? Which consonants are we transitioning into?</li> </ul>
46/59 Language and Computers Topic 1: Text and Speech Encoding	Sampling rate	47/59 Language and Computers Topic 1: Text and Speech Encoding	Applications of speech encoding
Writing systems Appatent Appatent Systems Appatent Systems Appatent Appaten	<ul> <li>The sampling rate is often 8000 or 16,000 samples per second. The rate for CDs is 44,100 samples/second (or Hertz (Hz))</li> <li>The higher the sampling rate, the better quality the recording but the more space it takes.</li> <li>Speech needs at least 8000 samples/second, but most likely 16,000 or 22,050 Hz will be used nowadays.</li> </ul>	Writing systems Aphabetic Systems with unsual realization Relation to language Comparison of systems Encoding written language AGCII Unicode Tanacrytion Tanacrytion Tanacrytion Tanacrytion Way geach is hard to represent Actuation Macaute Actuation Macaute Actuation	Mapping sounds to symbols (alphabet), and vice versa, isn't all that easy. • Automatic Speech Recognition (ASR): sounds to text • Text-to-Speech Synthesis (TTS): texts to sounds
49/59 Language and Computers Topic 1: Text and Speech Encoding	Kinds of ASR systems	50/59 Language and Computers Topic 1: Text and Speech Encoding	Kinds of ASR systems
Writing systems Aphabatic System Lapopsphile Lapopsphile Lapopsphile Assessments Systems Comparison d systems Encoding written Language Comparison d systems Comparison d systems Comparison Spoken Lando Spoken Lando Spoken Lando Mity speech Lando Articulatio Relating written and spoken Language	<ul> <li>Different kinds of systems:</li> <li>Speaker dependent = work for a single speaker</li> <li>Speaker independent = work for any speaker of a given variety of a language, e.g. American English</li> <li>Speaker adaptive = start as independent but begin to adapt to a single speaker to improve accuracy</li> </ul>	Writing systems Aphabatic System Lapopapitic Uncodentification Relation to language Comparison of systems Encoding written language Ascil Uncode Tamacrigion Tamacrigion Way speech is hard to represent Actuation Acoustics Relating written and spoken language Previous feet	<ul> <li>Differing sizes of vocabularies, from tens of words to tens of thousands of words</li> <li>continuous speech vs. isolated-word systems:         <ul> <li>continuous speech systems = words connected together and not separated by pauses</li> <li>isolated-word systems = single words recognized at a time, requiring pauses to be inserted between words</li> <li>→ easier to find the endpoints of words</li> </ul> </li> </ul>

52/59

From Text to Speer

From Text to Speech

Language and

Computers Topic 1: Text and Speech Encoding

Writing systems Alphabeti Syllabic Logographic Systems with unusual realization Relation to language Comparison of system Encoding written language ASCII Unicode Typing it in Spoken language

Why speech is hard to represent Articulation Acoustics

Relating written and

48/59

Language and

Computers

Topic 1: Text and Speech Encoding

Writing systems

Logographic Systems with unusual realization

Relation to language Come

Encoding written

Spoken language Why speech is hard to

Relating written and spoken language

Language and Computers

Topic 1: Text and Speech Encoding

Writing systems Alphabetic Syllabic Logographic

Systems with un realization

Relation to language Comparison of sy

Encoding written language

Why speech is hard to represent

Relating written and spoken language

From Text to Speech

ASCII Unicode

Typing it in Spoken language

Articulation Acoustics 51/59

From Speech to Text From Text to Speech

language ASCII Unicode Typing it in

Articulation Acoustics

Alphabeti

spoken language From Speech to Text From Text to Speech

# Steps in an ASR system

- 1. Digital sampling of speech
- 2. Acoustic signal processing = converting the speech samples into particular measurable units
- 3. Recognition of sounds, groups of sounds, and words

May or may not use more sophisticated analysis of the utterance to help.

# Speech to Text to Speech

If we convert speech to text and then back to speech, it should sound the same, right?

- But at the conversion stages, there is information loss. To avoid this loss would require a lot of memory and knowledge about what exact information to store.
- The process is thus irreversible.

# Text-to-Speech Synthesis (TTS)

Could just record a voice saying phrases or words and then play back those words in the appropriate order. Or can break the text down into smaller units

- 1. Convert input text into phonetic alphabet
- 2. Synthesize phonetic characters into speech
- To synthesize characters into speech, people have tried:
- using formulas which adjust the values of the frequencies, the loudness, etc.
- using a model of the vocal tract and trying to produce sounds based on how a human would speak

#### 55/59 Language and Computers Demos

Language and

Computers

Topic 1: Text and

Speech Encoding

Writing systems

Systems with unusual realization

Relation to language Comparison of sys

Encoding written

Spoken language

Why speech is hard to represent

Relating written and

Topic 1: Text and

Alphabeti Svllabic

Logographic Systems with unusual

Relation to language

Encoding written

Spoken language

language

ASCII

Unicode

Typing it in

spoken language

From Speech to Text

From Text to Speech

Alphabetic Syllabic

Logographic

language

ASCII

Unicode

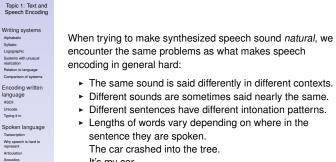
Typing it in

Articulation Acoustics

### Speech Encoding Writing systems

### Text-to-Speech

- AT&T mulitilingual TTS system: http://www.research.att.com/projects/tts/demo.php
- Nuance Realspeak: http://www.nuance.com/realspeak/demo/default.asp
- various systems and languages: http://www.ims.uni-stuttgart.de/~moehler/synthspeech/



Relating written and

Language and Computers

Topic 1: Text and Speech Encoding

Alphabi Syllabic

Comparison of s

language ASCII

Linicode

Typing it in

Articulation

Acoustics

56/59

spoken language From Speech to Text From Text to Speech

Language and It's hard to be natural Computers

It's my car.

The same sound is said differently in different contexts.

Different sounds are sometimes said nearly the same.

Different sentences have different intonation patterns.

Lengths of words vary depending on where in the

sentence they are spoken.

The car crashed into the tree.

Cars, trucks, and bikes are vehicles.

Language and Computers Topic 1: Text and Speech Encoding

Writing systems Alphabeti Syllabic Logographic Systems with unusual realization Relation to language Comparison of system Encoding written language ASCII Unicode Typing it in

Spoken language Why speech is hard to represent Articulation Acoustics

Relating written and spoken language From Speech to Text

57/59

# Writing systems Logographic Systems with unusual realization Relation to language Encoding written Spoken language Why speech is hard to represent Relating written and spoken language From Speech to Text From Text to Speech

59/59

Why speech is hard to represent Articulation Acoustics Relating written and spoken language From Speech to Text From Text to Speech

58/59