

# Linguistics 384: Language and Computers

## Topic 1: Text and Speech Encoding

Scott Martin\*

Dept. of Linguistics, OSU  
Spring 2008

### Writing systems

- Alphabetic
- Syllabic
- Logographic
- Systems with unusual realization
- Relation to language
- Comparison of systems

### 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

---

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

# Language and Computers – where to start?

- ▶ If we want to do anything with language, we need a way to represent language.

## Writing systems

- Alphabetic
- Syllabic
- Logographic
- Systems with unusual realization
- Relation to language
- Comparison of systems

## 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

# 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

## Writing systems

Alphabetic  
Syllabic  
Logographic  
Systems with unusual realization  
Relation to language  
Comparison of systems

## 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

# 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

## Writing systems

Alphabetic  
Syllabic  
Logographic  
Systems with unusual  
realization  
Relation to language  
Comparison of systems

## 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

## Writing systems

### Writing systems

- Alphabetic
- Syllabic
- Logographic
- Systems with unusual realization
- Relation to language
- Comparison of systems

### 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

## Writing systems

## Encoding written language

### Writing systems

- Alphabetic
- Syllabic
- Logographic
- Systems with unusual realization
- Relation to language
- Comparison of systems

### 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

Writing systems

Encoding written language

Spoken language

## Writing systems

- Alphabetic
- Syllabic
- Logographic
- Systems with unusual realization
- Relation to language
- Comparison of systems

## 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

Writing systems

Encoding written language

Spoken language

Relating written and spoken language

## Writing systems

- Alphabetic
- Syllabic
- Logographic
- Systems with unusual realization
- Relation to language
- Comparison of systems

## 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

## 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)*

### Writing systems

- Alphabetic
- Syllabic
- Logographic
- Systems with unusual realization
- Relation to language
- Comparison of systems

### 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

## 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>.

### Writing systems

- Alphabetic
- Syllabic
- Logographic
- Systems with unusual realization
- Relation to language
- Comparison of systems

### 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

## Alphabets (phonemic alphabets)

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

### Writing systems

#### Alphabetic

Syllabic

Logographic

Systems with unusual  
realization

Relation to language

Comparison of systems

### 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

## 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

### Writing systems

#### Alphabetic

##### Syllabic

##### Logographic

##### Systems with unusual realization

##### Relation to language

##### Comparison of systems

### 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

# Alphabet example: Fraser

An alphabet used to write Lisu, a Tibeto-Burman language spoken by about 657,000 people in Myanmar, India, Thailand and in the Chinese provinces of Yunnan and Sichuan.

## Consonants

P	ɸ	B	ɸ	W	M	M	T	ɬ	D	S	R	N	L	F	F
[p]	[pʰ]	[b]	[f]	[v]	[m]	[ɰ]	[t]	[tʰ]	[d]	[s]	[z]	[n]	[l]	[ts]	[tʂʰ]
Z	C	C	J	X	R	Y	K	K	G	H	B	Λ	G	V	
[dz]	[c]	[cʰ]	[ɟ]	[ʃ]	[ʒ]	[l]	[k]	[kʰ]	[g]	[x]	[ɣ]	[ŋ]	[ɦ]	[h]	

## Vowels

I	E	ʌ	ʊ	ɛ	ɾ	D	A	U	O
[i]	[e]	[æ]	[ü]	[ø]	[w]	[ə]	[ɑ]	[u]	[ɔ]

## Tones

.	,	..	..	:	:	'
high tone	mid rising	mid tone	mid tense	low tone	low tense	nasalization

(from: <http://www.omniglot.com/writing/fraser.htm>)

## Writing systems

### Alphabetic

### Syllabic

### Logographic

### Systems with unusual realization

### Relation to language

### Comparison of systems

## 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

# 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.

𐤀 ḥēt ḥ	𐤁 zayin z	𐤂 wāw w	𐤃 hē h	𐤄 dālet d	𐤅 gīmel g	𐤆 bēt b	𐤇 'ālef '
𐤈 sāmek s	𐤉 nun n	𐤊 mēm m	𐤋 lāmed l	𐤌 kaf k	𐤍 yōd y	𐤎 ṭēt ṭ	
𐤏 tāw t	𐤐 śin/śin ś	𐤑 rēš r	𐤒 qōf q	𐤓 šādē š	𐤔 pē p	𐤕 'ayin '	

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

## Writing systems

- Alphabetic
- Syllabic
- Logographic
- Systems with unusual realization
- Relation to language
- Comparison of systems

## 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

# A note on the letter-sound correspondence

- ▶ Alphabets use letters to encode sounds (consonants, vowels).

## Writing systems

Alphabetic

Syllabic

Logographic

Systems with unusual  
realization

Relation to language

Comparison of systems

## 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

# A note on the letter-sound correspondence

- ▶ Alphabets use letters to encode sounds (consonants, vowels).
- ▶ But the correspondence between spelling and pronunciation in many languages is quite complex, i.e., not a simple one-to-one correspondence.

## Writing systems

Alphabetic

Syllabic

Logographic

Systems with unusual  
realization

Relation to language

Comparison of systems

## 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

# A note on the letter-sound correspondence

- ▶ Alphabets use letters to encode sounds (consonants, vowels).
- ▶ But the correspondence between spelling and pronunciation in many languages is quite complex, i.e., not a simple one-to-one correspondence.
- ▶ Example: English

## Writing systems

### Alphabetic

Syllabic

Logographic

Systems with unusual  
realization

Relation to language

Comparison of systems

## 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

# A note on the letter-sound correspondence

- ▶ Alphabets use letters to encode sounds (consonants, vowels).
- ▶ But the correspondence between spelling and pronunciation in many languages is quite complex, i.e., not a simple one-to-one correspondence.
- ▶ Example: English
  - ▶ same spelling – different sounds: *ought*, *cough*, *tough*, *through*, *though*, *hiccough*

## Writing systems

Alphabetic

Syllabic

Logographic

Systems with unusual  
realization

Relation to language

Comparison of systems

## 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

# A note on the letter-sound correspondence

- ▶ Alphabets use letters to encode sounds (consonants, vowels).
- ▶ But the correspondence between spelling and pronunciation in many languages is quite 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*, *mortg**age*

## Writing systems

### Alphabetic

#### Syllabic

#### Logographic

#### Systems with unusual realization

#### Relation to language

#### Comparison of systems

## 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

# A note on the letter-sound correspondence

- ▶ Alphabets use letters to encode sounds (consonants, vowels).
- ▶ But the correspondence between spelling and pronunciation in many languages is quite 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*

## Writing systems

### Alphabetic

#### Syllabic

#### Logographic

Systems with unusual  
realization

Relation to language

Comparison of systems

## 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

# A note on the letter-sound correspondence

- ▶ Alphabets use letters to encode sounds (consonants, vowels).
- ▶ But the correspondence between spelling and pronunciation in many languages is quite 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*

## Writing systems

### Alphabetic

Syllabic

Logographic

Systems with unusual  
realization

Relation to language

Comparison of systems

## 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

# A note on the letter-sound correspondence

- ▶ Alphabets use letters to encode sounds (consonants, vowels).
- ▶ But the correspondence between spelling and pronunciation in many languages is quite 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*)

## Writing systems

### Alphabetic

Syllabic

Logographic

Systems with unusual  
realization

Relation to language

Comparison of systems

## 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

# More examples for non-transparent letter-sound correspondences

## French

- (1) a. *Versailles* → [ver*sai*]  
b. *ete, etais, etait, etaient* → [ete]

### Writing systems

#### Alphabetic

#### Syllabic

#### Logographic

#### Systems with unusual realization

#### Relation to language

#### Comparison of systems

### 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

# More examples for non-transparent letter-sound correspondences

## French

- (1) a. *Versailles* → [ver*sai*]  
b. *ete, etais, etait, etaient* → [ete]

## Irish

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

### Writing systems

Alphabetic

Syllabic

Logographic

Systems with unusual  
realization

Relation to language

Comparison of systems

### 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

## Writing systems

Alphabetic

Syllabic

Logographic

Systems with unusual  
realization

Relation to language

Comparison of systems

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

# More examples for non-transparent letter-sound correspondences

## French

- (1) a. *Versailles* → [ver*sai*]  
b. *ete, etais, etait, etaient* → [ete]

## Irish

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

What is the notation used within the []?

# The International Phonetic Alphabet (IPA)

- ▶ Several special alphabets for representing sounds have been developed, the best known being the International Phonetic Alphabet (IPA).

## Writing systems

### Alphabetic

Syllabic

Logographic

Systems with unusual  
realization

Relation to language

Comparison of systems

## 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

# 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.

## Writing systems

### Alphabetic

Syllabic

Logographic

Systems with unusual  
realization

Relation to language

Comparison of systems

## 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

# 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/IPA/lab/IPA/lab.htm>

## Writing systems

### Alphabetic

Syllabic

Logographic

Systems with unusual  
realization

Relation to language

Comparison of systems

## 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

# Syllabic systems

## 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>)

### Writing systems

Alphabetic

Syllabic

Logographic

Systems with unusual  
realization

Relation to language

Comparison of systems

### 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

# Syllabary example: Cypriote

The Cypriot syllabary or Cypro-Minoan writing is thought to have developed from the Linear A, or possibly the Linear B script of Crete, though its exact origins are not known. It was used from about 800 to 200 BC.

a	ta	ga	ka	pa	la	ma	na	ra	sa	ya	xa	ya
e	te		ke	pe	le	me	ne	re	se	ve	xe	
i	ti		ki	pi	li	mi	ni	ri	si	vi		
o	to		ko	po	bo	mo	no	ro	so	vo	zo	yo
u	tu		ku	pu	lu	mu	nu	ru	su	yu		

(from: <http://www.omniglot.com/writing/cypriot.htm>)

## Writing systems

Alphabetic

Syllabic

Logographic

Systems with unusual  
realization

Relation to language

Comparison of systems

## 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

# 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:

ກະ	ກິ	ກຸ	ກຸ່	ກາ	ກີ	ກູ	ກູ່	ເກະ	ເກະ
ka	ki	ku	ku'	ka:	ki:	ku:	ku:'	ke	kae
[ka]	[ki]	[ku]	[ku]	[ka:]	[ki:]	[ku:]	[ku:]	[ke]	[kae]
ໄກະ	ເກ	ເກ	ໄກ	ເກະ	ເກີ	ເກັວ	ເກຢ	ກົວ	ເກັວ
ko	ke:	kae:	ko:	ko'	koe	kia	kia	kua	koe:y
[ko]	[ke:]	[kae]	[ko:]	[kɔ]	[kɤ]	[kia]	[kia]	[kuə]	[kɤ:j]
ເກຢ	ກໍ	ເກີ	ເກືອ	ເກົາ	ໄກ	ໄກ	ກົ່	ກໍ	
koe:y	ko':	koe:	ku'a	kaw	kay	kay	kam	k	
[kɤ:j]	[kɔ:]	[kɤ:]	[kuə]	[kaw]	[kaj]	[kaj]	[kam]	[k]	

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

## Writing systems

Alphabetic

Syllabic

Logographic

Systems with unusual  
realization

Relation to language

Comparison of systems

## 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

# Logographic writing systems

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

## Writing systems

Alphabetic

Syllabic

Logographic

Systems with unusual  
realization

Relation to language

Comparison of systems

## 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

# Logographic writing systems

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

Example: development of Chinese character *horse*:



## Writing systems

Alphabetic

Syllabic

Logographic

Systems with unusual  
realization

Relation to language

Comparison of systems

## 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

# 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

## Writing systems

Alphabetic

Syllabic

Logographic

Systems with unusual realization

Relation to language

Comparison of systems

## 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







# Logograph writing system example: Chinese

## Pictographs

女子口日月山川豕目心雨田木龜  
woman child mouth sun moon mountain river pig eye heart rain field tree turtle

### Writing systems

Alphabetic

Syllabic

#### Logographic

Systems with unusual  
realization

Relation to language

Comparison of systems

### 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



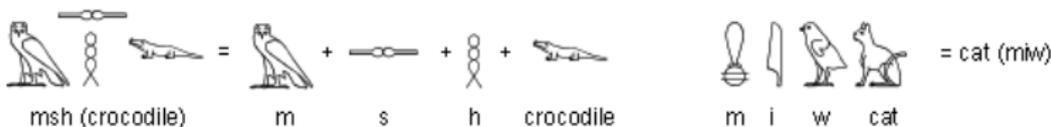




# Semantic-phonetic compounds

	phonetic component					
	古 gǔ	扁 biǎn	敖 áo	旁 páng	堯 yáo	
semantic component (radical)	人 (person)	估 gū (to guess)	偏 piān (biased)	傲 ào (proud)	傍 bāng (beside)	僥 jiǎo (lucky)
	言 (words)	話 huà (commentaries)	辯 biàn (to quibble)	謗 wàng (to slander)	謗 bàng (to libel)	饒 riào (to argue)
	虫 (insect)	蛄 gū (mole cricket)	蝙 biān (bat)	螯 áo ([crab's] nippers)	螃 páng (crab)	蟯 xiào (worm)
	金 (metal)	鈷 gǔ (cobalt)		鏊 áo (griddle)	鎊 bàng (pound sterling)	饒 riào (cymbals)

## An example from Ancient Egyptian



(from: <http://www.omniglot.com/writing/egyptian.htm>)

### Writing systems

Alphabetic

Syllabic

Logographic

Systems with unusual realization

Relation to language

Comparison of systems

### 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

# 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.

### Writing systems

Alphabetic  
Syllabic  
Logographic

#### Systems with unusual realization

Relation to language  
Comparison of systems

### 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

## 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](http://www.library.cornell.edu/africana/Writing_Systems/Chroma.html))

# Braille alphabet

•	••	•••	••••	••	•••	••••	•••	••	•••	•	••	•••
A	B	C	D	E	F	G	H	I	J	K	L	M
1	2	3	4	5	6	7	8	9	0	knowledge	like	more
•••	••	•••	••••	•••	••	•••	••	•••	•••	•••	•••	•••
N	O	P	Q	R	S	T	U	V	W	X	Y	Z
not		people	quite	rather	so	that	us	very	will	it	you	as
•••	••••	•••	••	•••	•	••	•••	•••	••	•••	•••	•••
Ç	É	À	È	Ù	Â	Ê	Î	Ô	Û	Ë	Ï	Ü
and	for	of	the	with	child ch	gh	shall sh	this th	which wh	ed	er	out ou
•••	•	••	•••	•••	••	•••	•••	•••	••	•••	••	••
Ö œ	,	;	:	.		!	()	? "	*	"	fraction line st	Ò
ow		bb	cc	dd	en		gg; were		in			ing
•••	•••	•	••	•	••	••	••	•				
numeral sign	Ä Æ	'	-	numerical index accent	literal index	italic sign decimal sign	letter sign	capital sign				

## Writing systems

Alphabetic  
Syllabic  
Logographic

Systems with unusual  
realization

Relation to language  
Comparison of systems

## 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

# Chromatographic system



## Writing systems

- Alphabetic
- Syllabic
- Logographic

## Systems with unusual realization

- Relation to language
- Comparison of systems

## 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

# 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).

## Writing systems

Alphabetic

Syllabic

Logographic

Systems with unusual  
realization

## Relation to language

Comparison of systems

## 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

# 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

## Writing systems

Alphabetic

Syllabic

Logographic

Systems with unusual  
realization

## Relation to language

Comparison of systems

## 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

Japanese: logographic system *kanji*, syllabary *katakana*,  
syllabary *hiragana*

- ▶ *kanji*: 5,000-10,000 borrowed Chinese characters

## Writing systems

- Alphabetic
- Syllabic
- Logographic
- Systems with unusual realization

## Relation to language

- Comparison of systems

## 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

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

## Writing systems

Alphabetic

Syllabic

Logographic

Systems with unusual  
realization

## Relation to language

Comparison of systems

## 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

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

## Writing systems

Alphabetic

Syllabic

Logographic

Systems with unusual  
realization

## Relation to language

Comparison of systems

## 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

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

## Writing systems

Alphabetic

Syllabic

Logographic

Systems with unusual  
realization

## Relation to language

Comparison of systems

## 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

# Japanese example

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

## カプセルホテル

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/japanese.htm#origin>)

### Writing systems

Alphabetic

Syllabic

Logographic

Systems with unusual  
realization

### Relation to language

Comparison of systems

### 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

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

## Writing systems

- Alphabetic
- Syllabic
- Logographic
- Systems with unusual realization

## Relation to language

- Comparison of systems

## 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

“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)

## Writing systems

Alphabetic  
Syllabic  
Logographic  
Systems with unusual realization

## Relation to language

Comparison of systems

## 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

“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.

## Writing systems

Alphabetic  
Syllabic  
Logographic  
Systems with unusual realization

## Relation to language

Comparison of systems

## 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

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

- ▶ 7th century until 1920s: Arabic scripts. Three different Arabic scripts used

## Writing systems

Alphabetic  
Syllabic  
Logographic  
Systems with unusual realization

## Relation to language

Comparison of systems

## 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

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.

## Writing systems

Alphabetic  
Syllabic  
Logographic  
Systems with unusual realization

## Relation to language

Comparison of systems

## 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

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

## Writing systems

Alphabetic  
Syllabic  
Logographic  
Systems with unusual realization

## Relation to language

Comparison of systems

## 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

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

## Writing systems

Alphabetic  
Syllabic  
Logographic  
Systems with unusual realization

## Relation to language

Comparison of systems

## 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

# Comparison of writing systems

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

## Writing systems

Alphabetic

Syllabic

Logographic

Systems with unusual  
realization

Relation to language

Comparison of systems

## 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

# Comparison of writing systems

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

- ▶ accuracy: Can every word be written down accurately?

## Writing systems

Alphabetic

Syllabic

Logographic

Systems with unusual  
realization

Relation to language

Comparison of systems

## 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

# Comparison of writing systems

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

- ▶ accuracy: Can every word be written down accurately?
- ▶ learnability: How long does it take to learn the system?

## Writing systems

Alphabetic

Syllabic

Logographic

Systems with unusual  
realization

Relation to language

Comparison of systems

## 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

# Comparison of writing systems

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

- ▶ accuracy: Can every word be written down accurately?
- ▶ learnability: How long does it take to learn the system?
- ▶ cognitive ability: Are some systems unnatural? (e.g. Does dyslexia show that alphabets are unnatural?)

## Writing systems

Alphabetic

Syllabic

Logographic

Systems with unusual  
realization

Relation to language

Comparison of systems

## 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

# Comparison of writing systems

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

- ▶ accuracy: Can every word be written down accurately?
- ▶ learnability: How long does it take to learn the system?
- ▶ cognitive ability: Are some systems unnatural? (e.g. Does dyslexia show that alphabets are unnatural?)
- ▶ language-particular differences: English has thousands of possible syllables; Japanese has very few in comparison

## Writing systems

Alphabetic

Syllabic

Logographic

Systems with unusual  
realization

Relation to language

Comparison of systems

## 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

# Comparison of writing systems

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

- ▶ accuracy: Can every word be written down accurately?
- ▶ learnability: How long does it take to learn the system?
- ▶ cognitive ability: Are some systems unnatural? (e.g. Does dyslexia show that alphabets are unnatural?)
- ▶ language-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?

## Writing systems

Alphabetic  
Syllabic  
Logographic  
Systems with unusual realization  
Relation to language  
Comparison of systems

## 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

# Encoding written language

- ▶ Information on a computer is stored in **bits**.
- ▶ A bit is either on (= 1, yes) or off (= 0, no).

## Writing systems

Alphabetic  
Syllabic  
Logographic  
Systems with unusual  
realization  
Relation to language  
Comparison of systems

## 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

# 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 1
    - ▶ “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

## Writing systems

Alphabetic

Syllabic

Logographic

Systems with unusual  
realization

Relation to language

Comparison of systems

## 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

# 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
?	?	?	?

## Writing systems

- Alphabetic
- Syllabic
- Logographic
- Systems with unusual realization
- Relation to language
- Comparison of systems

## 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

# 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	?	?	?

## Writing systems

- Alphabetic
- Syllabic
- Logographic
- Systems with unusual realization
- Relation to language
- Comparison of systems

## 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

# 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.

## Writing systems

- Alphabetic
- Syllabic
- Logographic
- Systems with unusual realization
- Relation to language
- Comparison of systems

## 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

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

# 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.

## Writing systems

- Alphabetic
- Syllabic
- Logographic
- Systems with unusual realization
- Relation to language
- Comparison of systems

## 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

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

# 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.

## Writing systems

- Alphabetic
- Syllabic
- Logographic
- Systems with unusual realization
- Relation to language
- Comparison of systems

## 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

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

# Converting decimal numbers to binary - Division Method

Decimal	Remainder?	Binary
$10/2 = 5$	no	0

## Writing systems

- Alphabetic
- Syllabic
- Logographic
- Systems with unusual realization
- Relation to language
- Comparison of systems

## 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

# Converting decimal numbers to binary - Division Method

Decimal	Remainder?	Binary
$10/2 = 5$	no	0
$5/2 = 2$	yes	10

## Writing systems

- Alphabetic
- Syllabic
- Logographic
- Systems with unusual realization
- Relation to language
- Comparison of systems

## 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

# Converting decimal numbers to binary - Division Method

Decimal	Remainder?	Binary
$10/2 = 5$	no	0
$5/2 = 2$	yes	10
$2/2 = 1$	no	010

## Writing systems

- Alphabetic
- Syllabic
- Logographic
- Systems with unusual realization
- Relation to language
- Comparison of systems

## 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

# Converting decimal numbers to binary - Division Method

Decimal	Remainder?	Binary
$10/2 = 5$	no	0
$5/2 = 2$	yes	10
$2/2 = 1$	no	010
$1/2 = 0$	yes	1010

## Writing systems

- Alphabetic
- Syllabic
- Logographic
- Systems with unusual realization
- Relation to language
- Comparison of systems

## 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

# Converting decimal numbers to binary - Division Method

Decimal	Remainder?	Binary
$10/2 = 5$	no	0
$5/2 = 2$	yes	10
$2/2 = 1$	no	010
$1/2 = 0$	yes	1010

## Writing systems

- Alphabetic
- Syllabic
- Logographic
- Systems with unusual realization
- Relation to language
- Comparison of systems

## 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

# Using bytes to store characters

With 8 bits (a single byte), you can represent 256 different characters. Why would we want so many?

## Writing systems

- Alphabetic
- Syllabic
- Logographic
- Systems with unusual realization
- Relation to language
- Comparison of systems

## 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

# Using bytes to store characters

With 8 bits (a single byte), you can represent 256 different characters. Why would we want so many?

- ▶ If you look at a keyboard, you will find lots of non-English characters.
- ▶ 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.

## Writing systems

Alphabetic  
Syllabic  
Logographic  
Systems with unusual realization  
Relation to language  
Comparison of systems

## 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

# An encoding standard: ASCII

- ▶ **ASCII** = the American Standard Code for Information Interchange
- ▶ 7-bit code for storing English text
- ▶ 7 bits = 128 possible characters.
- ▶ The numeric order reflects alphabetic ordering.

## Writing systems

Alphabetic  
Syllabic  
Logographic  
Systems with unusual  
realization  
Relation to language  
Comparison of systems

## 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



- ▶ 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. ]

[Some characters may be displayed incorrectly. ]

## Writing systems

- Alphabetic
- Syllabic
- Logographic
- Systems with unusual realization
- Relation to language
- Comparison of systems

## 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

- ▶ 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. ]

[Some characters may be displayed incorrectly. ]

- ▶ 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.

## Writing systems

Alphabetic

Syllabic

Logographic

Systems with unusual  
realization

Relation to language

Comparison of systems

## 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

- ▶ 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. ]

[Some characters may be displayed incorrectly. ]

- ▶ 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.
- ▶ 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.

## Writing systems

Alphabetic

Syllabic

Logographic

Systems with unusual  
realization

Relation to language

Comparison of systems

## 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

# Multipurpose Internet Mail Extensions (MIME)

MIME provides meta-information on the text, which tells us:

## Writing systems

- Alphabetic
- Syllabic
- Logographic
- Systems with unusual realization
- Relation to language
- Comparison of systems

## 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

# Multipurpose Internet Mail Extensions (MIME)

MIME provides meta-information on the text, which tells us:

- ▶ which version of MIME is being used
- ▶ what the character set is
- ▶ if that character set was altered, how it was altered

Mime-Version: 1.0 Content-Type: text/plain; charset=US-ASCII

Content-Transfer-Encoding: 7bit

## Writing systems

Alphabetic  
Syllabic  
Logographic  
Systems with unusual  
realization  
Relation to language  
Comparison of systems

## 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

# Different coding systems

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

## Writing systems

- Alphabetic
- Syllabic
- Logographic
- Systems with unusual realization
- Relation to language
- Comparison of systems

## 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

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

## Writing systems

Alphabetic  
Syllabic  
Logographic  
Systems with unusual realization  
Relation to language  
Comparison of systems

## 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

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 → **Unicode**

## Writing systems

Alphabetic  
Syllabic  
Logographic  
Systems with unusual realization  
Relation to language  
Comparison of systems

## 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

## Problems with having multiple encoding systems:

- ▶ Conflicts: two encodings can use the same number for two different characters and use different numbers for the same character.
- ▶ Hassle: have to install many, many systems if you want to be able to deal with various languages

### Writing systems

Alphabetic  
Syllabic  
Logographic  
Systems with unusual realization  
Relation to language  
Comparison of systems

### 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

Problems with having multiple encoding systems:

- ▶ Conflicts: two encodings can use the same number for two different characters and use different numbers for the same character.
- ▶ Hassle: have to install many, many systems if you want to be able to deal with various languages

Unicode tries to fix that by having a single representation for every possible character.

*“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](http://www.unicode.org))*

## Writing systems

Alphabetic  
Syllabic  
Logographic  
Systems with unusual realization  
Relation to language  
Comparison of systems

## 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

# How big is Unicode?

Version 3.2 has codes for 95,221 characters from alphabets, syllabaries and logographic systems.

- ▶ Uses 32 bits – meaning we can store  $2^{32} = 4,294,967,296$  characters.
- ▶ 4 billion possibilities for each character? That takes a lot of space on the computer!

## Writing systems

Alphabetic  
Syllabic  
Logographic  
Systems with unusual  
realization  
Relation to language  
Comparison of systems

## 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):  $2^{16} = 65536$
  - ▶ UTF-8 (8 bits):  $2^8 = 256$
- ▶ How is it possible to encode  $2^{32}$  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.

## Writing systems

Alphabetic  
Syllabic  
Logographic  
Systems with unusual  
realization  
Relation to language  
Comparison of systems

## 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

# How do we type everything in?

- ▶ 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.

## Writing systems

Alphabetic  
Syllabic  
Logographic  
Systems with unusual realization  
Relation to language  
Comparison of systems

## 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

# How do we type everything in?

- ▶ 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.

## Writing systems

Alphabetic  
Syllabic  
Logographic  
Systems with unusual  
realization  
Relation to language  
Comparison of systems

## 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

# How do we type everything in?

- ▶ 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

## Writing systems

Alphabetic  
Syllabic  
Logographic  
Systems with unusual  
realization  
Relation to language  
Comparison of systems

## 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

# How do we type everything in?

- ▶ 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.

## Writing systems

Alphabetic  
Syllabic  
Logographic  
Systems with unusual  
realization  
Relation to language  
Comparison of systems

## 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

# Unwritten languages

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

## Writing systems

Alphabetic  
Syllabic  
Logographic  
Systems with unusual realization  
Relation to language  
Comparison of systems

## 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

# 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?

## Writing systems

Alphabetic  
Syllabic  
Logographic  
Systems with unusual  
realization  
Relation to language  
Comparison of systems

## 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

# 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

## Writing systems

Alphabetic  
Syllabic  
Logographic  
Systems with unusual  
realization  
Relation to language  
Comparison of systems

## 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

# What does speech look like?

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

## Writing systems

- Alphabetic
- Syllabic
- Logographic
- Systems with unusual realization
- Relation to language
- Comparison of systems

## 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

# 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.

## Writing systems

Alphabetic  
Syllabic  
Logographic  
Systems with unusual  
realization  
Relation to language  
Comparison of systems

## 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

# 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.
  - ⇒ We need to know:
    - ▶ some properties of speech
    - ▶ how to measure these speech properties
    - ▶ how these measurements correspond to sounds we hear

## Writing systems

Alphabetic  
Syllabic  
Logographic  
Systems with unusual realization  
Relation to language  
Comparison of systems

## 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

# What makes representing speech hard?

## Difficulties:

- ▶ People have different dialects and different size vocal tracts and thus say things differently

## Writing systems

Alphabetic  
Syllabic  
Logographic  
Systems with unusual  
realization  
Relation to language  
Comparison of systems

## 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

# 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.

### Writing systems

Alphabetic  
Syllabic  
Logographic  
Systems with unusual  
realization  
Relation to language  
Comparison of systems

### 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

# 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*.

## Writing systems

Alphabetic  
Syllabic  
Logographic  
Systems with unusual realization  
Relation to language  
Comparison of systems

## 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

# 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* in *see* is very acoustically similar to the *sh* in *shoe*

## Writing systems

Alphabetic  
Syllabic  
Logographic  
Systems with unusual realization  
Relation to language  
Comparison of systems

## 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

# Articulatory properties: How it's produced

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

## Writing systems

- Alphabetic
- Syllabic
- Logographic
- Systems with unusual realization
- Relation to language
- Comparison of systems

## 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

# 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]

## Writing systems

Alphabetic  
Syllabic  
Logographic  
Systems with unusual realization  
Relation to language  
Comparison of systems

## 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

# 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*.

## Writing systems

Alphabetic  
Syllabic  
Logographic  
Systems with unusual realization  
Relation to language  
Comparison of systems

## 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

# Acoustic properties: What it sounds like

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

⇒ Akin to ripples in a pond

- ▶ **speech flow** = rate of speaking, number and length of pauses (seconds)
- ▶ **loudness** (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

## Writing systems

Alphabetic  
Syllabic  
Logographic  
Systems with unusual realization  
Relation to language  
Comparison of systems

## Encoding written language

ASCII  
Unicode  
Typing it in

## Spoken language

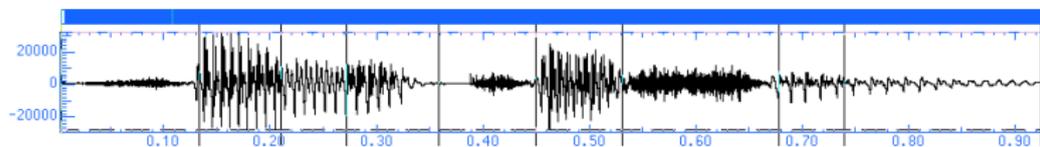
Transcription  
Why speech is hard to represent  
Articulation

## Acoustics

## Relating written and spoken language

From Speech to Text  
From Text to Speech

# Oscillogram (Waveform)



## Writing systems

- Alphabetic
- Syllabic
- Logographic
- Systems with unusual realization
- Relation to language
- Comparison of systems

## Encoding written language

- ASCII
- Unicode
- Typing it in

## Spoken language

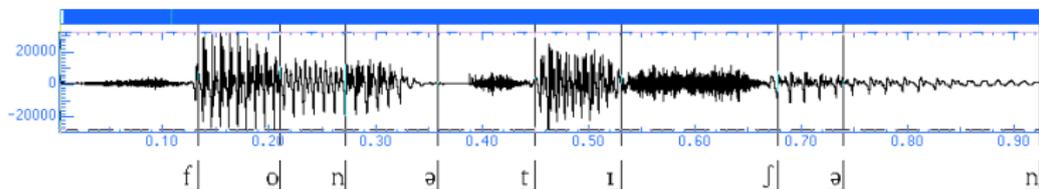
- Transcription
- Why speech is hard to represent
- Articulation

## Acoustics

## Relating written and spoken language

- From Speech to Text
- From Text to Speech

# Oscillogram (Waveform)



(Check out the *Speech Analysis Tutorial*, of the Department of Linguistics at Lund University, Sweden at <http://www.ling.lu.se/research/spechtutorial/tutorial.html>, from which the illustrations on this and the following slides are taken.)

## Writing systems

- Alphabetic
- Syllabic
- Logographic
- Systems with unusual realization
- Relation to language
- Comparison of systems

## Encoding written language

- ASCII
- Unicode
- Typing it in

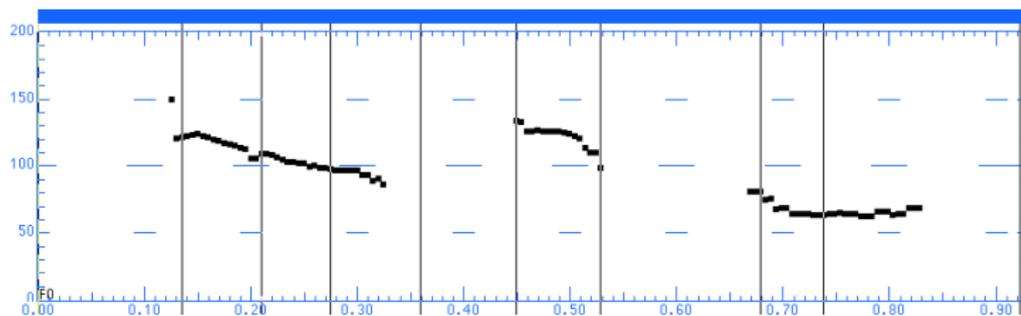
## Spoken language

- Transcription
- Why speech is hard to represent
- Articulation
- Acoustics

## Relating written and spoken language

- From Speech to Text
- From Text to Speech

# Fundamental frequency (F0, pitch)



## Writing systems

- Alphabetic
- Syllabic
- Logographic
- Systems with unusual realization
- Relation to language
- Comparison of systems

## Encoding written language

- ASCII
- Unicode
- Typing it in

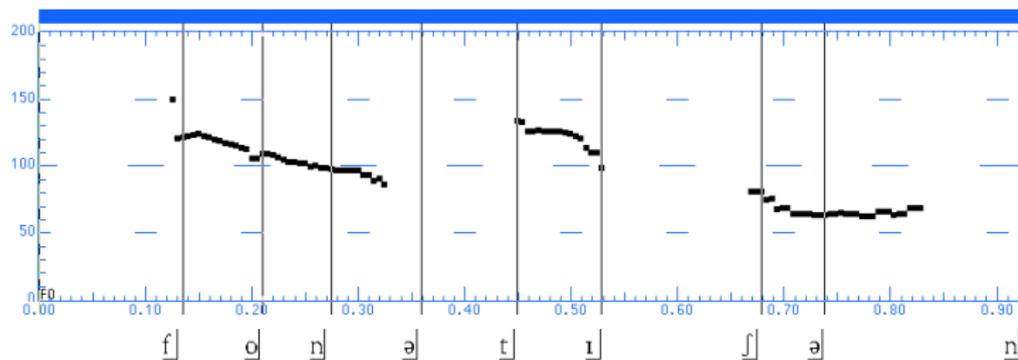
## Spoken language

- Transcription
- Why speech is hard to represent
- Articulation
- Acoustics

## Relating written and spoken language

- From Speech to Text
- From Text to Speech

# Fundamental frequency (F0, pitch)



## Writing systems

- Alphabetic
- Syllabic
- Logographic
- Systems with unusual realization
- Relation to language
- Comparison of systems

## Encoding written language

- ASCII
- Unicode
- Typing it in

## Spoken language

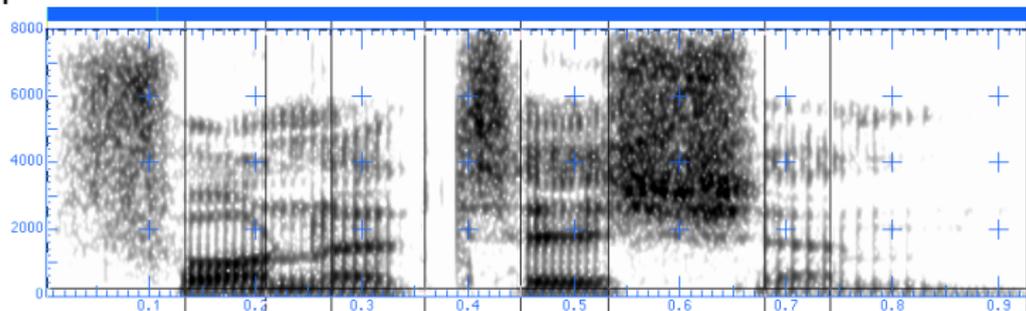
- Transcription
- Why speech is hard to represent
- Articulation
- Acoustics

## Relating written and spoken language

- From Speech to Text
- From Text to Speech

# Spectrograms

**Spectrogram** = a graph to represent (the frequencies of) speech over time.



## Writing systems

- Alphabetic
- Syllabic
- Logographic
- Systems with unusual realization
- Relation to language
- Comparison of systems

## Encoding written language

- ASCII
- Unicode
- Typing it in

## Spoken language

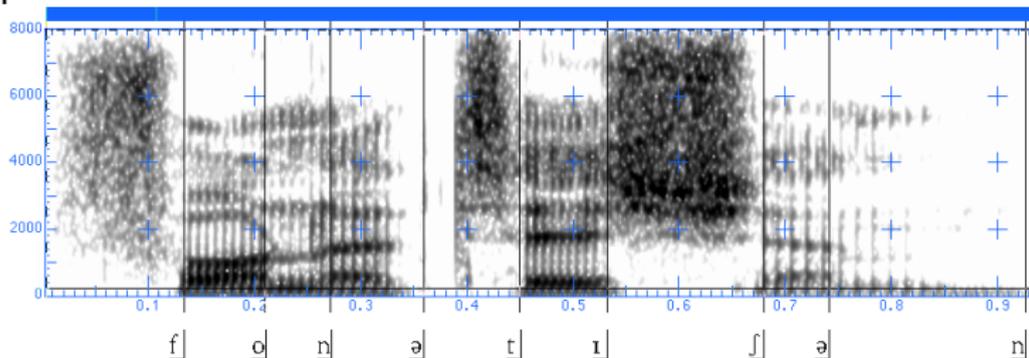
- Transcription
- Why speech is hard to represent
- Articulation
- Acoustics

## Relating written and spoken language

- From Speech to Text
- From Text to Speech

# Spectrograms

**Spectrogram** = a graph to represent (the frequencies of) speech over time.



## Writing systems

- Alphabetic
- Syllabic
- Logographic
- Systems with unusual realization
- Relation to language
- Comparison of systems

## 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

# How measurements correspond to sounds we hear

- ▶ How dark is the picture? → How loud is the sound?  
We can measure this in decibels.

## Writing systems

- Alphabetic
- Syllabic
- Logographic
- Systems with unusual realization
- Relation to language
- Comparison of systems

## 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

# How measurements correspond to sounds we hear

- ▶ How dark is the picture? → How loud is the sound?  
We can measure this in decibels.
- ▶ 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.

## Writing systems

Alphabetic  
Syllabic  
Logographic  
Systems with unusual realization  
Relation to language  
Comparison of systems

## 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

# How measurements correspond to sounds we hear

- ▶ How dark is the picture? → How loud is the sound?  
We can measure this in decibels.
- ▶ 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.
- ▶ How do these dark lines change? → How are the frequencies changing over time?  
Which consonants are we transitioning into?

## Writing systems

Alphabetic  
Syllabic  
Logographic  
Systems with unusual realization  
Relation to language  
Comparison of systems

## 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

# How did we get these measurements?

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

## Writing systems

- Alphabetic
- Syllabic
- Logographic
- Systems with unusual realization
- Relation to language
- Comparison of systems

## 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

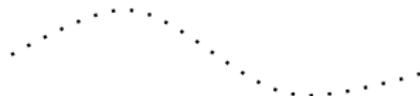
# How did we get these measurements?

**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.



CONTINUOUS



DISCRETE

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

## Writing systems

- Alphabetic
- Syllabic
- Logographic
- Systems with unusual realization
- Relation to language
- Comparison of systems

## 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

# Sampling rate

- ▶ The sampling rate is often 8000 or 16,000 samples per second. The rate for CDs is 44,100 samples/second (or **Hertz (Hz)**)

## Writing systems

- Alphabetic
- Syllabic
- Logographic
- Systems with unusual realization
- Relation to language
- Comparison of systems

## 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

# Sampling rate

- ▶ The sampling rate is often 8000 or 16,000 samples per second. The rate for CDs is 44,100 samples/second (or **Hertz (Hz)**)
- ▶ The higher the sampling rate, the better quality the recording ... but the more space it takes.

## Writing systems

- Alphabetic
- Syllabic
- Logographic
- Systems with unusual realization
- Relation to language
- Comparison of systems

## 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

# Sampling rate

- ▶ The sampling rate is often 8000 or 16,000 samples per second. The rate for CDs is 44,100 samples/second (or **Hertz (Hz)**)
- ▶ The higher the sampling rate, the better quality the recording ... but the more space it takes.
- ▶ Speech needs at least 8000 samples/second, but most likely 16,000 or 22,050 Hz will be used nowadays.

## Writing systems

Alphabetic  
Syllabic  
Logographic  
Systems with unusual  
realization  
Relation to language  
Comparison of systems

## 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

# Applications of speech encoding

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

## Writing systems

Alphabetic  
Syllabic  
Logographic  
Systems with unusual realization  
Relation to language  
Comparison of systems

## 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

# Automatic Speech Recognition (ASR)

Automatic speech recognition = process by which the computer maps a speech signal to text.

## Writing systems

- Alphabetic
- Syllabic
- Logographic
- Systems with unusual realization
- Relation to language
- Comparison of systems

## 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

# 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

## Writing systems

Alphabetic

Syllabic

Logographic

Systems with unusual  
realization

Relation to language

Comparison of systems

## 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

# Kinds of ASR systems

Different kinds of systems:

- ▶ Speaker dependent = work for a single speaker
- ▶ Speaker independent = work for any speaker of a given variety of a language, e.g. American English
- ▶ Speaker adaptive = start as independent but begin to adapt to a single speaker to improve accuracy

## Writing systems

Alphabetic  
Syllabic  
Logographic  
Systems with unusual  
realization  
Relation to language  
Comparison of systems

## 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

# Kinds of ASR systems

- ▶ Differing sizes of vocabularies, from tens of words to tens of thousands of words

## Writing systems

Alphabetic  
Syllabic  
Logographic  
Systems with unusual  
realization  
Relation to language  
Comparison of systems

## 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

# Kinds of ASR systems

- ▶ Differing sizes of vocabularies, from tens of words to tens of thousands of words
- ▶ **continuous speech** vs. **isolated-word** systems:
  - ▶ continuous speech systems = words connected together and not separated by pauses
  - ▶ isolated-word systems = single words recognized at a time, requiring pauses to be inserted between words  
→ easier to find the endpoints of words

## Writing systems

Alphabetic  
Syllabic  
Logographic  
Systems with unusual  
realization  
Relation to language  
Comparison of systems

## 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

# Steps in an ASR system

## 1. Digital sampling of speech

### Writing systems

- Alphabetic
- Syllabic
- Logographic
- Systems with unusual realization
- Relation to language
- Comparison of systems

### 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

# Steps in an ASR system

1. Digital sampling of speech
2. **Acoustic signal processing** = converting the speech samples into particular measurable units

## Writing systems

- Alphabetic
- Syllabic
- Logographic
- Systems with unusual realization
- Relation to language
- Comparison of systems

## 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

# 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

## Writing systems

- Alphabetic
- Syllabic
- Logographic
- Systems with unusual realization
- Relation to language
- Comparison of systems

## 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

# 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.

## Writing systems

- Alphabetic
- Syllabic
- Logographic
- Systems with unusual realization
- Relation to language
- Comparison of systems

## 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

# 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

## Writing systems

- Alphabetic
- Syllabic
- Logographic
- Systems with unusual realization
- Relation to language
- Comparison of systems

## 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

# 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

## Writing systems

Alphabetic  
Syllabic  
Logographic  
Systems with unusual realization  
Relation to language  
Comparison of systems

## 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

# It's hard to be natural

When trying to make synthesized speech sound *natural*, we encounter the same problems as what makes speech encoding in general hard:

- ▶ 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.

It's my car.

Cars, trucks, and bikes are vehicles.

## Writing systems

Alphabetic  
Syllabic  
Logographic  
Systems with unusual  
realization  
Relation to language  
Comparison of systems

## 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

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**.

## Writing systems

Alphabetic  
Syllabic  
Logographic  
Systems with unusual realization  
Relation to language  
Comparison of systems

## 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

## Text-to-Speech

- ▶ AT&T multilingual 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/>

### Writing systems

Alphabetic  
Syllabic  
Logographic  
Systems with unusual  
realization  
Relation to language  
Comparison of systems

### 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