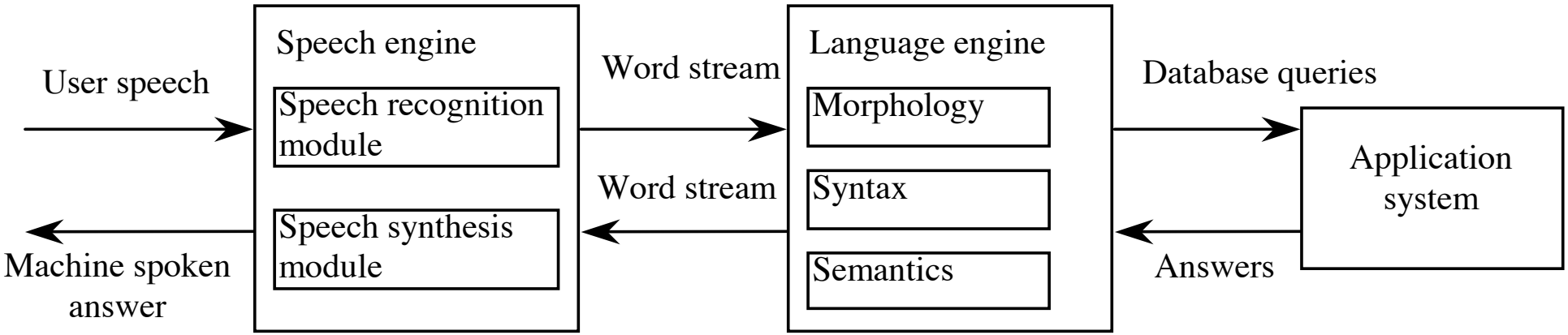
Language Technology
http://cs.lth.se/edan20/
Chapter 18: Speech Synthesis

Pierre Nugues

Pierre.Nugues@cs.lth.se
http://cs.lth.se/pierre_nugues/

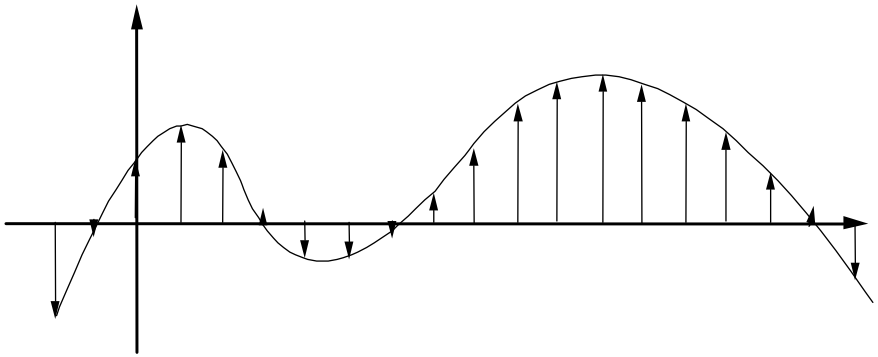
October 15, 2020

Structure of a Spoken Interactive System

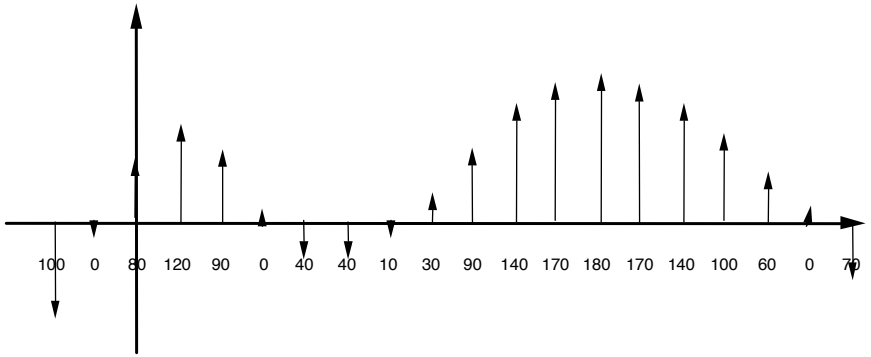


Signals

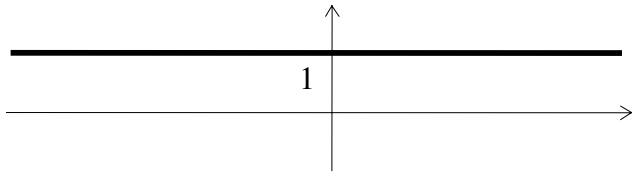
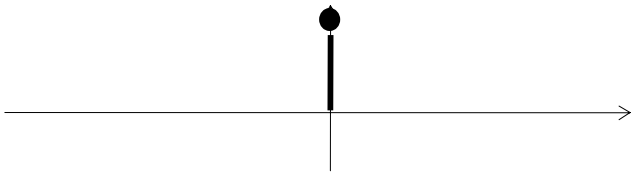
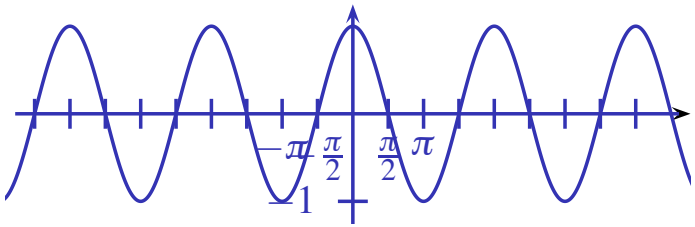
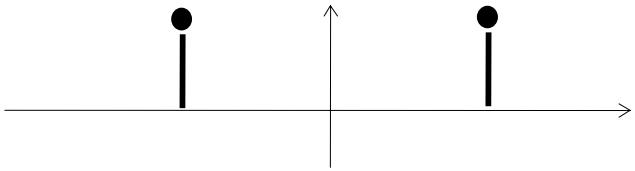
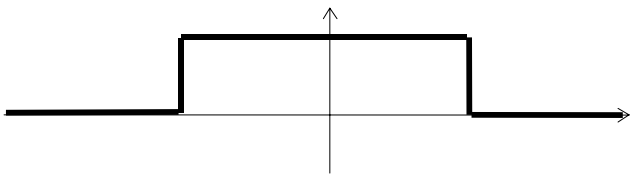
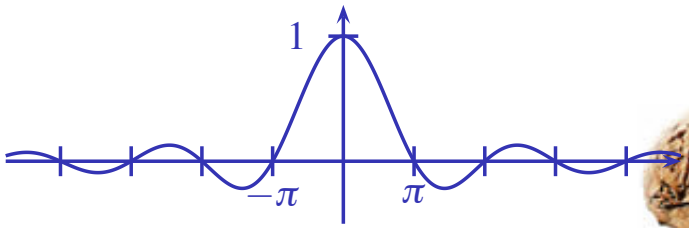
Sampling



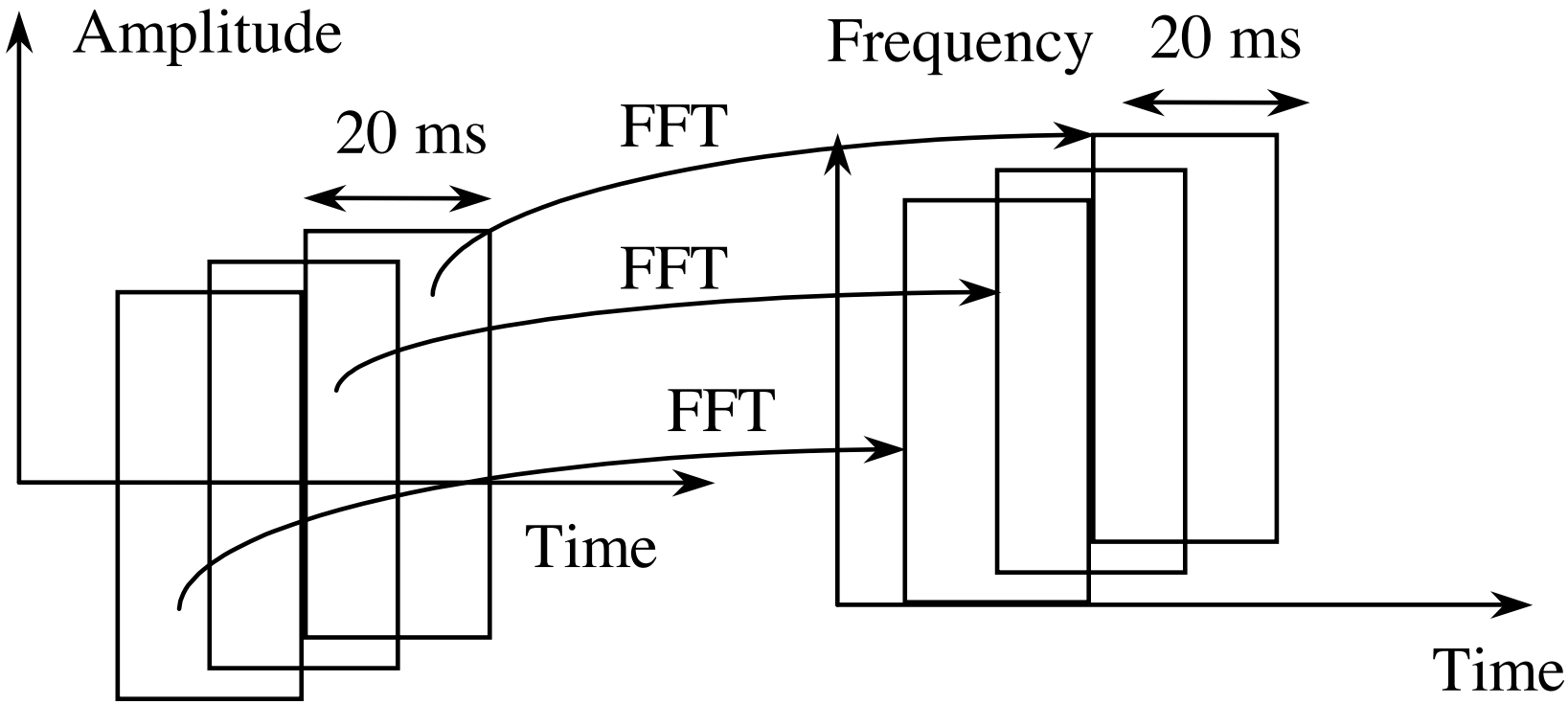
Digitization



Fourier Transforms

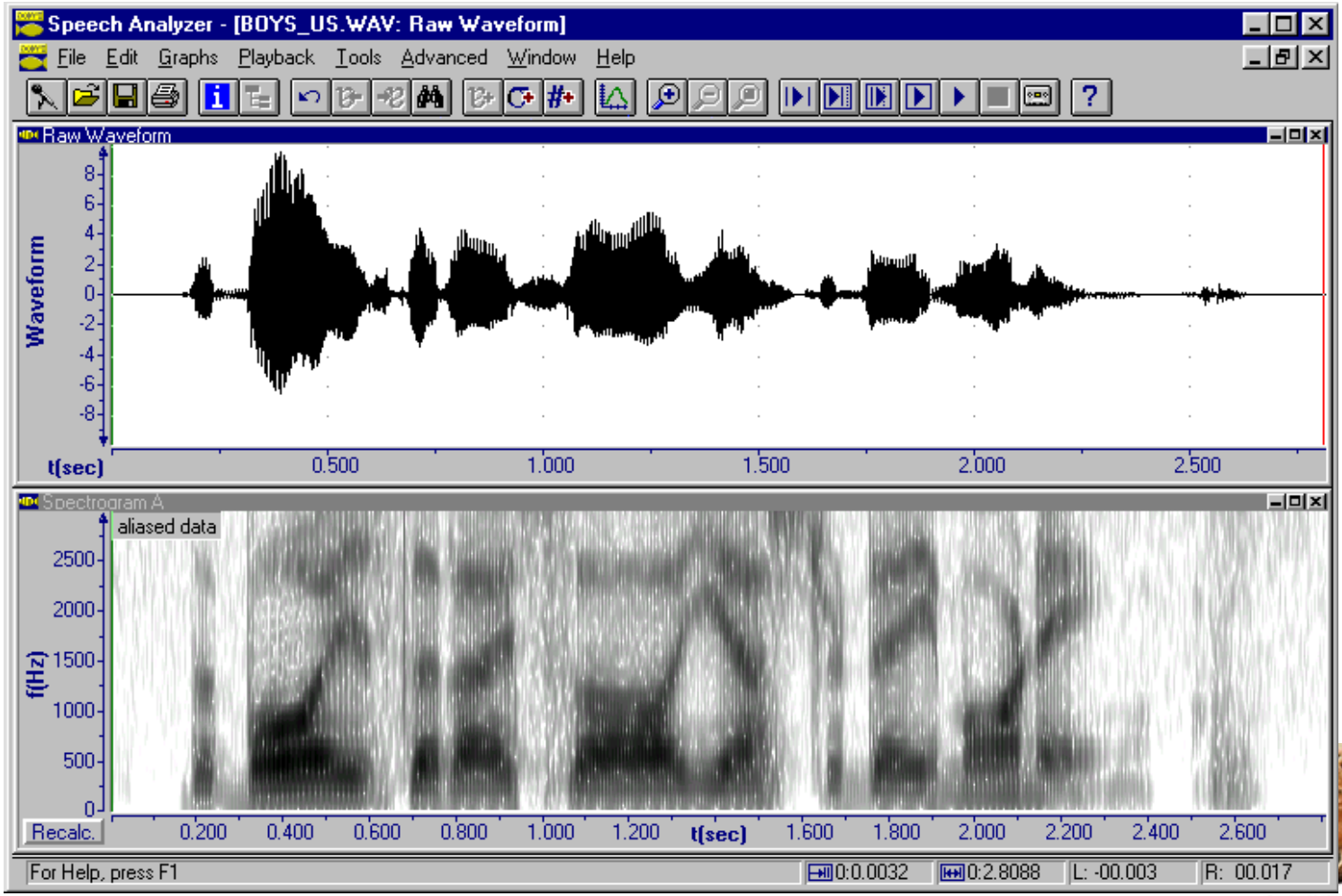
Time domain	**Frequency domain (Fourier Transforms)**
Unit constant function: $f(x) = 1$	Delta function, perfect impulse at 0: $\delta(x)$
	
Cosine: $\cos(2\pi\omega x)$	Shifted deltas: $\frac{\delta(x+\omega)+\delta(x-\omega)}{2}$
	
Square pulse : $w_a(x) = \begin{cases} 1 & -\frac{1}{2} \leq x \leq \frac{1}{2} \\ 0 & elsewhere \end{cases}$	$\mathrm{sinc}(x) = \frac{\sin(\pi x)}{\pi x}$
	

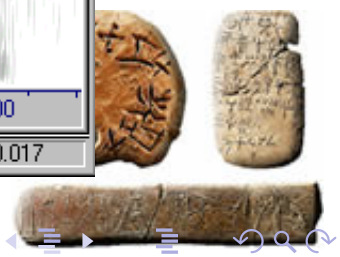
Speech Spectrograms



Speech Signals

The boys I saw yesterday morning





Phonemes

Phonemes are conceptual units to delimit elementary speech segments.
A broad phonemic transcription is denoted between slashes /symbol/
Phones are real speech sounds
Allophones are the members of the phone collection represented by a same phoneme.
Allophones can sometimes be predicted by the articulation context.
A narrow phonemic transcription is denoted between square angles [transcription]
Phonemes are divided into vowels and consonants

The IPA Notation

A notation to transcribe phonemes and allophones
Each language has a finite set of phonemes, around 40-60.

- Swedish has 18 consonants and 17 vowels
- French has 18 consonants, 14 vowels, and 3 semi-vowels (approximants)
- English has 24 consonants and 15 vowels.

Phonemes are specific to a language: *true* and *trou* 'hole' have the same broad transcription /tru/ but the narrow transcription is different [t̪ɹ̥u] and [t̪ʁ̥u]

Vowels

Vowels are voiced (F0) and have typical formant values: F1, F2, and F3. In North American English:

Formants (Hz)	/iː/	/ɪ/	/ɛ/	/æ/	/ɑ/	/ɔ/
F1	270	390	530	660	730	570
F2	2290	1990	1840	1720	1090	840
F3	3010	2550	2480	2410	2440	2410

The vowels can be classified according to the tongue position in the mouth.

Consonants

Consonants obstruct the airflow. They can be voiced or not.
They are classified using two parameters: the place and the manner of obstruction.

	Labial	Labio-dental	Dental	Alveolar	Post-alveolar	Palatal	Velar	Glottal
Plosive	p b			t d			k g	ʔ
Affricate					tʃ dʒ			
Nasal	m			n			ŋ	
Fricative		f v	θ ð	s z	ʃ ʒ			h
Approximant				r		j	w	
Lateral approximant					l			

Manner of Articulation

- Plosives block the oral cavity for a short period and release the air.
- Nasals let the air flow in the nasal cavity while blocking the oral cavity
- Fricatives restrict the airflow
- Approximants are vowel-like consonants: voiced and with little obstruction

Suprasegmental Features

A suprasegmental feature is a characteristic that extends over more than one phoneme or is independent of it as the stress that applies to a syllable.
The pitch, loudness, and quantity are amongst the most notable suprasegmental features.
They correspond to physical properties, respectively the fundamental frequency, the intensity (or amplitude), and the duration.
The relation between physical and perceptual properties is not trivial however.

Speech Synthesis

Use pre-recorded messages (train stations airports)
Use pre-recorded segments (phrases, words)
Map phonemes onto sound units → does not work well because of co-articulation
Two main techniques:

- Formant synthesis that works like an electronic music synthesizer
- Diphones concatenation that uses pre-recorded sound units.

The second method is generally better.
But phonemes don't transcribe directly to phones

Grapheme-to-Phoneme Conversion

Letters don't always map to a single phoneme as *give* and *life*
The conversion of graphemes into phonemes consists of:

- Tokenization.
- Dictionary lookup to process the exceptions.
- Morphological rules should be applied and may be irregular: *played* and *worked*, but *rugged* and *ragged*.
- Use of rules to process the rest of words supposed to be regular → right and left contexts of a grapheme

The venerable DECtalk has a lexicon of 7,000 words and 500 rules

Transcription Rules

The transcription rule format is similar to what we saw with morphological processing.

```
X --> y / <lc> _ <rc>
```

Rules may have no constraint on their left or right context as the rules

```
X --> y / _ <rc>
X --> y / <lc> _
```

or be context-free as the rule

```
X --> y
```

An Example

A simplified model of the pronunciation of the letter *c* in English is either /s/ before *e*, *i*, or *y* or /k/ elsewhere. The rules governing the transcription are

```
c --> s / _ {e, i, y}
c --> k/ _ {a, b, c, d, f, g, h, j, k, l, m, n, o, p,
                    q, r, s, t, u, v, w, x, z, #}
```

The transcription rules can be implemented with a transducer just like for morphology

POS Tagging

POS

I use *(verb) and* **a use** *(noun),*
to object *(verb) and* **an object** *(noun).*

French adverbs →

chantent *and* **notamment**

Semantics:

You get your just deserts
In the desert of Sudan.

Phone Concatenation

Use a database of prerecorded diphones, 3-phones, up to 5-phones
Segment *Paris* /pæris/ and use the diphone sequence:
#P, PA, AR, RI, IS, and S#.
Adjust suprasegmental parameters: the phone duration, intensity, and fundamental frequency (pitch value)

Phone Concatenation (II)

Diphones		Duration	Intensity	Pitch
#P	#[p]	70	80	120
PA	[pæ]	100	80	180
AR	[ær]	100	70	140
RI	[rɪ]	70	70	120
IS	[ɪs]	70	60	100
S#	[s]#	70	60	80

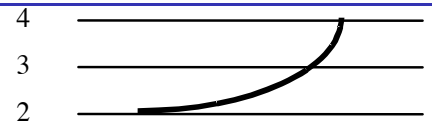
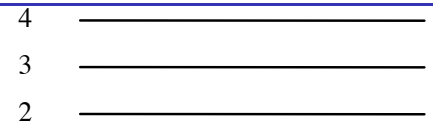
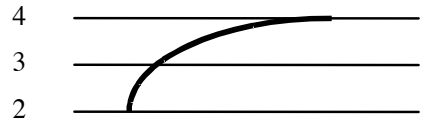
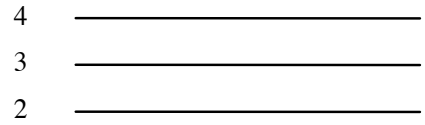
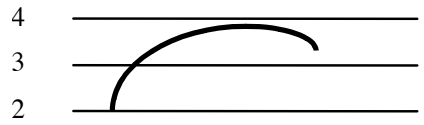
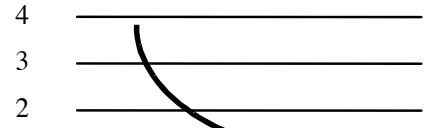
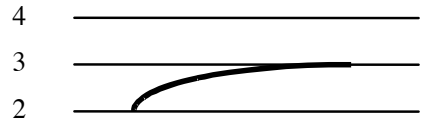
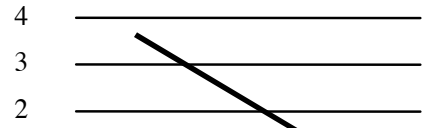
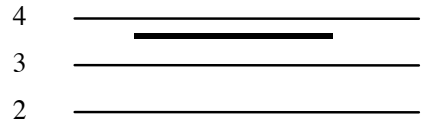
Prosody

Prosody corresponds to the melody and rhythm of speech.
It conveys syntactic, semantic as well as emotional information.
Prosodic aspects are often divided into features such as in English stress and intonation.
It applies differently to questions and declarations:

- *Yes/no* questions such as *Is it correct?*
- Other questions such as *What do you want?*

Prosody is implemented by adjusting intensity, duration, and pitch parameters

Intonation in French

Type	Pitch pattern	Type	Pitch pattern
Question (yes/no)		Parenthesis	
Major continuation		Finality	
Implication		*Wh*-question	
Minor continuation		Order	
Echo		Exclamation