2021

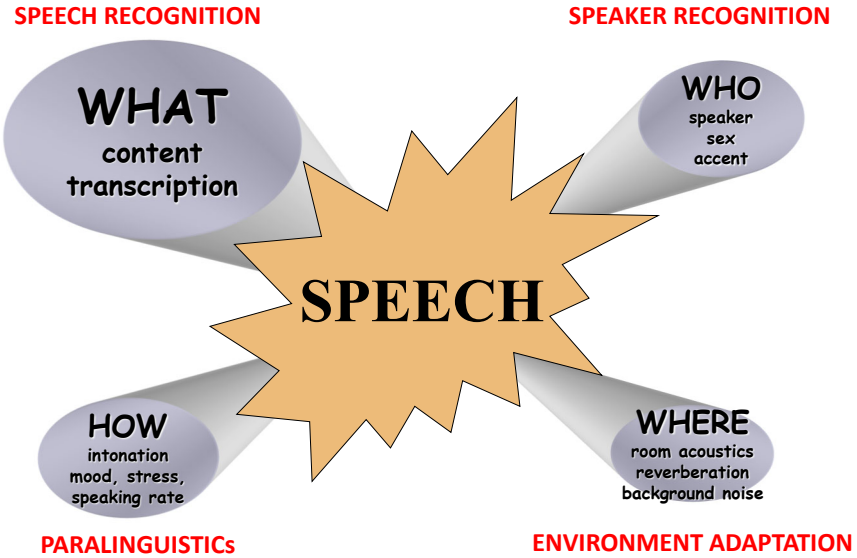
Speech Recognition (H02A6)

Dirk Van Compernolle

CHAPTER 0

Introduction

Speech Recognition Goal = Extracting Information from Spoken Language



Speech Recognition -- What is it ?

"The style of written prose in not that of spoken oratory, … "
Aristotle, Rethoric, Book 3 – part 12, 350 BC.

- **The goal of Speech Recognition can have subtle, though important, differences:**
  - **Text creation**
    … that generates a polished transcription that makes abstraction of irrelevant information in the input speech such as hesitations, flawed pronunciations, …
    … and is smart enough to add unspoken punctuation, understands markup commands, …
  - **Literal transcription**
    … as you want to see in the transcript of a session in court
  - **Speech understanding**
    ... Where only the content of underlying message matters, as needed for a voice search on the web or when operating equipment by voice

- **It's a scientific field full if ambiguities, as:**
  - **there exists no one-to-one mapping from acoustics to speech sounds**
  - **the speech signal is continuous whereas written language is discrete**

Speech Recognition by Humans and Machines

"airplanes don't flap wings "
The Book of Wisdom, 1st millennium BC.

- ***"Inspired"* by Human Speech Recognition**
  - **Machinery = human brain**
    - Crafted by millions of years of evolution
    - Trainable , Language independent
  - **Many years of Person & Language Dependent Learning / Adaptation**

Listening ➔ **babble** ➔ **sounds** ➔ **words** ➔ **sentences**





- ***"Executed"* by Machine Learning suitable for:**
  - ***Continuous-to-discrete* pattern matching**
  - ***Sequence-to-sequence* recognition**

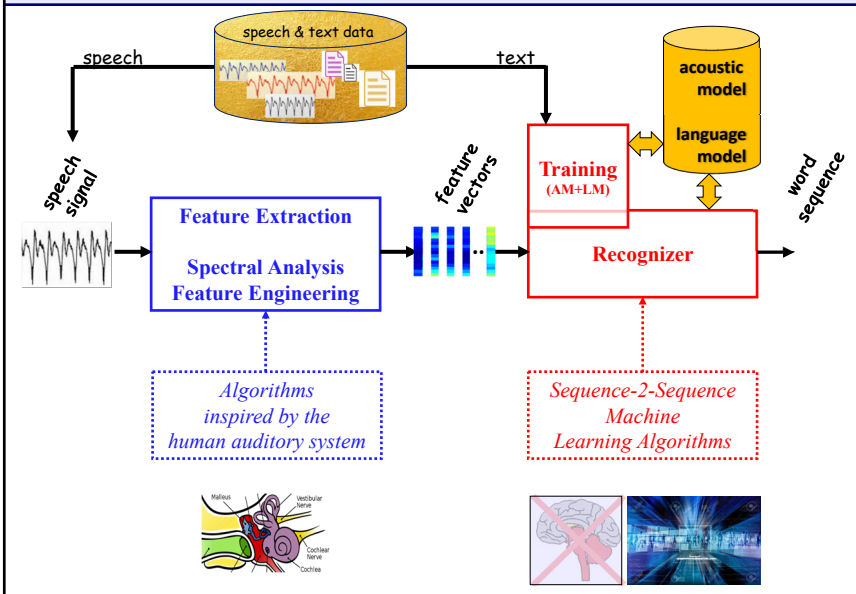
➔ **DTW, HMMs, DNNs, VITERBI, CTC, ….**

Large *speech + text* Databases

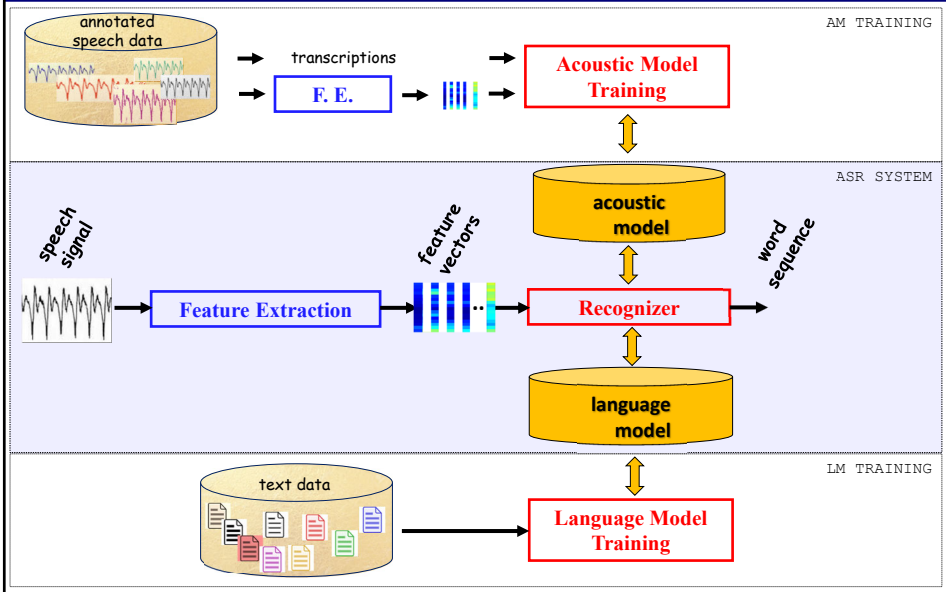
Acoustic Model

Language Model

Speech Recognition Architecture

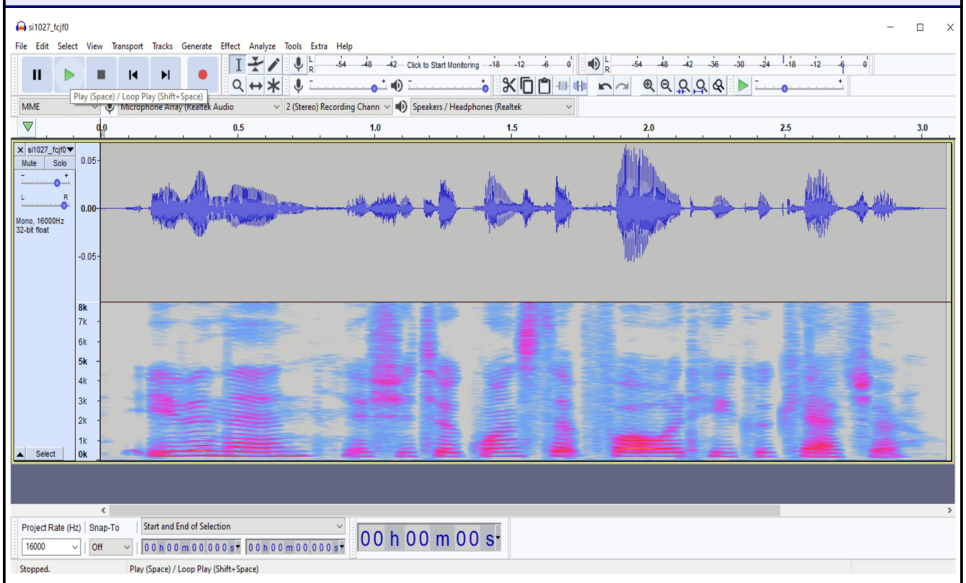


Speech Recognition Architecture

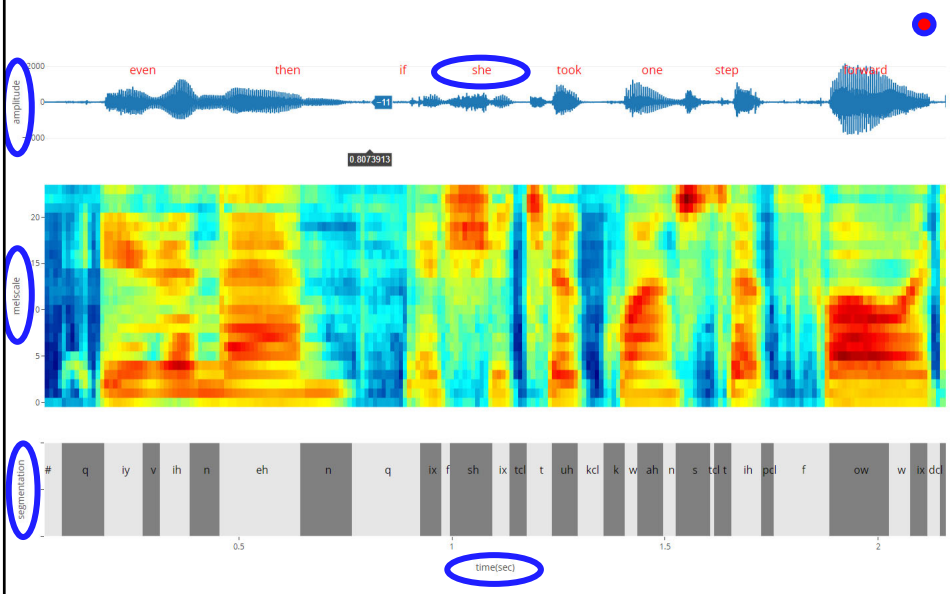


LOOKING at SPEECH and AUDIO

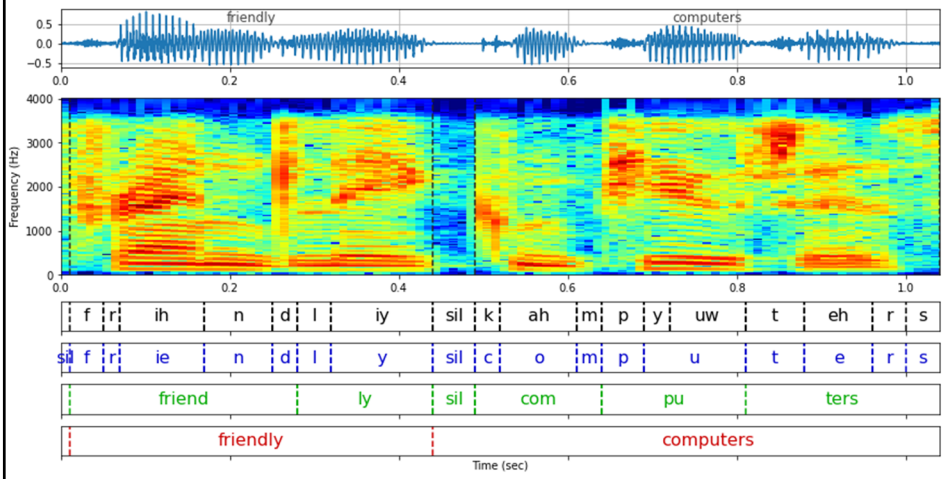
Tools for audio visualization / editing :
- Audacity: freeware, predecessor of Audition
- Adobe Audition (Adobe creative cloud)



LOOKING at SPEECH (ASR perspective)



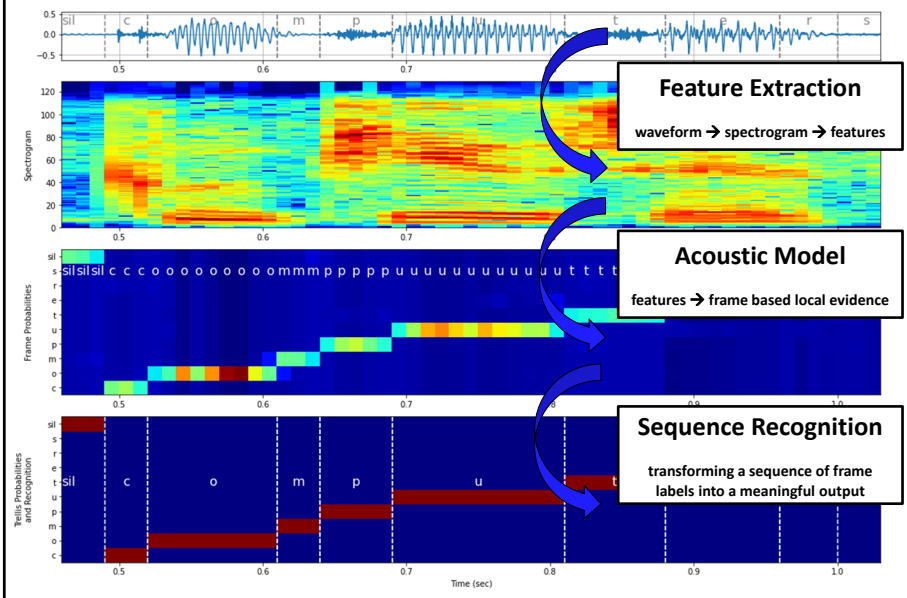
Transcribing Speech



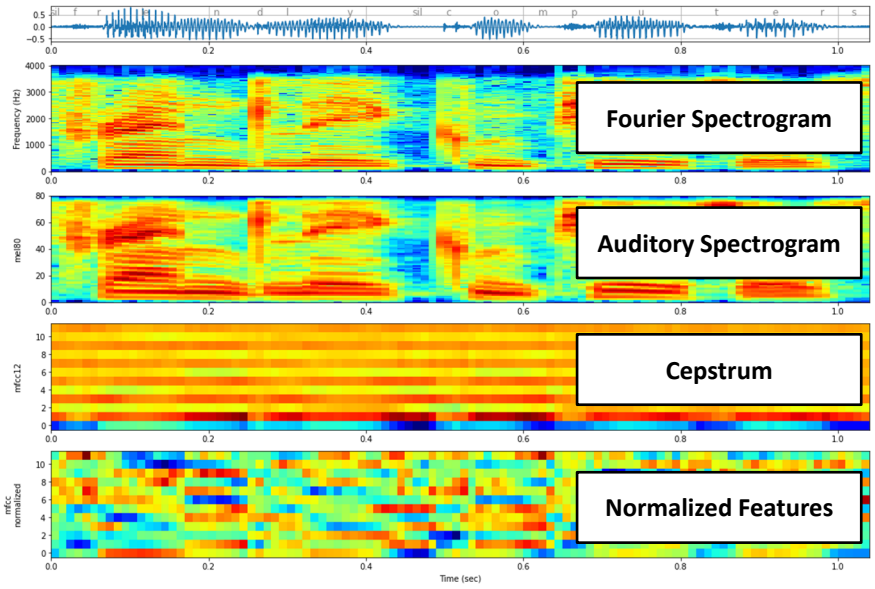
transcriptions and segmentations at the **phoneme**, **character**, **syllable** or **word** level

WARNING: ambiguity increases with level of detail !!!

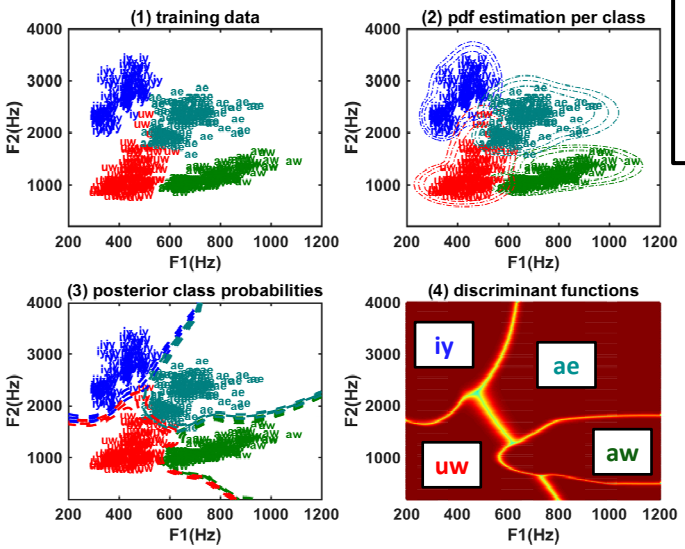
Speech Recognition in a nutshell



Feature Extraction (example)



ACOUSTIC MODEL (example)
Pattern matching of feature vectors



INPUT
feature vector

ACOUSTIC MODEL
= computational model of the input-output relationship
= machine learning: trained from example data, using HMM/GMMs, HMM/DNNs, RNNs, ..

OUTPUT
label scores

Speech Recognition in a Bayesian Framework

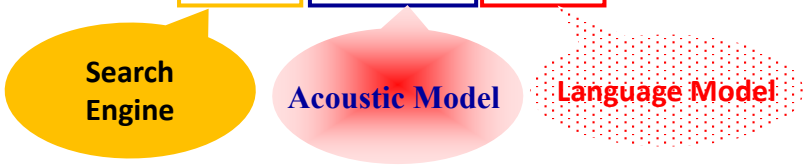
$X_{1:T}$ = feature vector sequence

W = word sequence

$$W^* = \underset{W}{\operatorname{argmax}} \; p(W \mid X_{1:T})$$

$$= \underset{W}{\operatorname{argmax}} \; \frac{p(X_{1:T} \mid W)\,p(W)}{\cancel{p(X_{1:T})}}$$

$$= \underset{W}{\operatorname{argmax}} \; p(X_{1:T} \mid W)\; p(W)$$



SPEECH RECOGNIZER

The speech recognizer (search engine, decoder)

1. **Converts an audio stream into a sequence of feature vectors (eg. one every 10 msec)**

2. **finds *(search for)* the most likely word sequence**
   - given the acoustic evidence in the acoustic model
   - given the linguistic evidence in the language model

Linguistic Components
support the acoustic core of a recognizer to resolve intrinsic ambiguities and improve recognition quality in general

Linguistic Resources (1): LEXICON

```
ABOUT        AH B AW T
ACTION       AE K SH AH N
CHEESE       CH IY Z
COMPUTER     K AH M P Y UW T ER
DESERT       D EH Z ER T
DESERT(1)    D IH Z ER T
DIVE         D AY V
GOOD         G UH D
HIDDEN       HH IH D AH N
JOY          JH OY
LANGUAGE     L AE NG G W AH JH
LANGUAGE(1)  L AE NG G W IH JH
LEARNING     L ER N IH NG
MACHINE      M AH SH IY N
MARKOV       M AA R K AO F
MODEL        M AA D AH L
MOTHER       M AH DH ER
NETWORKS     N EH T W ER K S
NEURAL       N UH R AH L
NEURAL(1)    N Y UH R AH L
PLEASURE     P L EH ZH ER
RECOGNITION  R EH K AH G N IH SH AH N
SHOUT        SH AW T
SPEECH       S P IY CH
SPOON        S P UW N
THING        TH IH NG
TOMATO       T AH M EY T OW
TOMATO(1)    T AH M AA T OW
TOMATOE      T AH M EY T OW
TOMATOE(1)   T AH M AA T OW
YES          Y EH S
```

- The lexicon gives the pronunciation of each word in the language
  - The Roman alphabet was not designed for the English language.
  - Due to different dynamics between written language (spelling reforms, loan words) and spoken language (uncontrolled natural evolution, dialects and accents) discrepancies exist between spoken and written forms
  - The lexicon includes pronunciation variants (tomato) and homonyms (desert)

Phonetic Alphabet

ARPABET	IPA	EXAMPLE	TRANSCRIPTION	PHONE CLASS
AA	ɑ	balm	B AA L M	VOWEL
AE	æ	bat	B AE T	VOWEL
AH	ʌ	butt	B AH T	VOWEL
AO	ɔ	bought	B AO T	VOWEL
AW	aʊ	bout	B AW T	VOWEL
AY	aɪ	bite	B AY T	VOWEL
B	b	buy	B AY	STOP
CH	tʃ	church	CH ER CH	AFFRICATE
D	d	die	D AY	STOP
DH	ð	thy	DH AY	FRICATIVE
EH	ɛ	bet	B EH T	VOWEL
ER	ɝ	bird	B ER T	VOWEL
EY	eɪ	bait	B EY T	VOWEL
F	f	fight	F AY T	FRICATIVE
G	g	guy	G AY	STOP
HH	h	high	HH AY	SEMIVOWEL
IH	ɪ	bit	B IH T	VOWEL
IY	i	beat	B IY T	VOWEL
JH	dʒ	jive	JH AY V	AFFRICATE
K	k	kite	K AY T	STOP
L	l	lie	L AY	LIQUID
M	m	my	M AY	NASAL
N	n	nigh	N AY	NASAL
NG	ŋ	sing	S IH NG	NASAL
OW	oʊ	boat	B OW T	VOWEL
OY	ɔɪ	boy	B OY	VOWEL
P	p	pie	P AY	STOP
R	ɹ	rye	R AY	LIQUID
S	s	sigh	S AY	FRICATIVE
SH	ʃ	shy	SH AY	FRICATIVE
T	t	tie	T AY	STOP
TH	θ	thigh	TH AY	FRICATIVE
UH	ʊ	book	B UH K	VOWEL
UW	u	boot	B UW T	VOWEL
V	v	vie	V AY	FRICATIVE
W	w	wise	W AY Z	SEMIVOWEL
Y	j	yacht	Y AH T	SEMIVOWEL
Z	z	zoo	Z UH	FRICATIVE
ZH	ʒ	pleasure	P L EH ZH ER	FRICATIVE

- IPA (International Phonetic Alphabet)
  - allows for a representation of spoken language
  - allows for different levels of detail
- ARPABET
  - is a simplified ASCII rendering of the IPA (International Phonetic Alphabet)
  - in the CMU Lexicon 39 phonetic symbols are used
  - optionally 3 levels of stress may be added to the vowels
  - additional symbols may be used to transcribe non-speech events: silence, coughs, hesitations, garbled speech, …
- Notes:
  - This is not a course on phonetics … we use our intuition in in these matters
  - The CMU LEXICON is a reference lexicon for US English used in many speech recognition benchmarks

Linguistic Resources (2): LANGUAGE MODEL

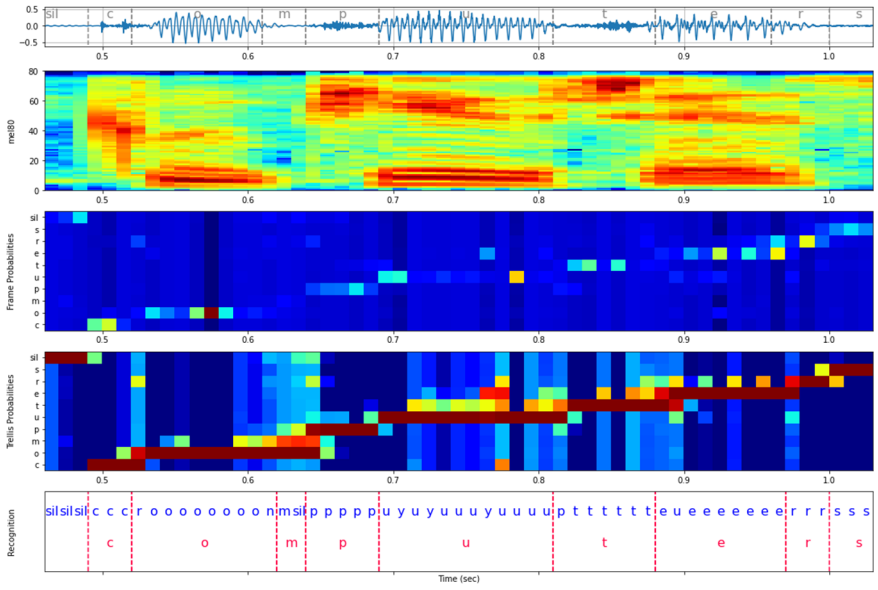
A statistical language model facilitates the computation of word sequence probabilities :

- **It implicitly knows about words: which words exist (lexicon) or how can they be formed (morphology)**
- **it allows for disambiguating homonyms on basis of the context**

  " twenty two people are too many to talk to "
- **It may drastically improve speech recognition accuracy by favoring probable word sequences over unlikely ones when the acoustic model is uncertain**

context	next word	probability
<s> I	want	.10
<s> I	would	.10
<s> I	don't	.03
...		
I want	to	.48
I want	a	.10
I want	that	.05
I want	you	.01
…		

Speech Recognition
it's not that easy … as there is a lot of noise on the line



Course Contents (1)

- PART I: SPEECH
  - Spectrogram: Time-Frequency Processing of Speech and Audio
  - Elementary Psycho-Acoustics: pitch, rhythm and timbre
  - Speech Generation with the Source-Filter Model and links to Acoustic-Phonetics
  - Exercises: Spectrogram, Auditory Demonstrations
- PART II: FRAME and SEGMENT RECOGNITION (CLASSIFICATION)
  - Statistical concepts and techniques for pattern matching
  - Speech features: pitch, formants, spectral envelope, mel-spectrum, cepstrum
  - Classifiers using acoustic-phonetic features (pitch, formants)
  - MLPs/DNNs for classification
  - Exercises: Frame Based Recognition with statistical techniques, Frame Based Recognition with DNNs

Course Contents (2)

- PART III: SPEECH RECOGNITION with SEQUENCE-2-SEQUENCE TECHNIQUES
  - Sequence Recognition
  - ASR Architecture
  - Hidden Markov Models
  - Context-Dependent Modeling with Decision Trees
  - Deep Neural Nets for sequence modeling (TDNN, CTC, LSTM)
  - Exercises: Dynamic Time Warping, Viterbi scoring and alignment, HMM Training, Decision Tree Design
- PART IV: ASR SYSTEMS
  - Linguistic Components: Lexicon & Language Models
  - Search
  - Deep Neural Nets for Language Modeling
  - L11: System Issues, Applications, Industrial Outlook
  - Exercises: Deep Neural Nets for LM

Practical Information H02A6

- **LECTURES**
  - **'live'-sessions are the reference**
  - **live streaming & recordings**
    - best effort, no guarantee
    - misses out on interaction & discussions
    - available via TOLEDO
- **EXERCISE SESSIONS**
  - **2 time slots per week with enough space (Tue. 13:30 and Wed. 16:00)**
    - choose which one fits best
  - **'live'-sessions are recommended**
  - **TA's are available DURING exercise time slots**
  - **All exercises can be run in the cloud**
  - **Addition Q&A sessions in COLLABORATE (TBD)**
- **PREREQUISITES:**
  - **cfr. Website, if in doubt do the selftest**
  - **computer skills:**
    - working with Jupyter Notebooks / Google Colab
    - Python programming: not more than changing a few parameters in simple code snippets
- **EXAM: open book, exercises**

Course Materials

- **WEBSITE: http://homes.esat.kuleuven.be/~compi/H02A6/**
  - **generic information on the course**
- **TOLEDO:**
  - **up-to-date scheduling info**
  - **all pointers to materials, recordings, exercise solutions, …**
- **GITHUB:**
  - **compi1234/pyspch: a Python package that is used throughout the course (exercises)**
  - **compi1234/spchlab: organized by chapter**
    - ch0
    - ch1 …
      - copies of lecture notes (html5)
      - .ipynb Jupyter notebooks used to create (many of the) figures in the course notes
      - .ipynb Jupyter notebooks with example code
      - .ipynb Jupyter notebooks used for the exercises
  - **MOSTLY NEW in 2021 and WORK IN PROGRESS (don't rely on materials related to lectures/exercises in the future) !!**