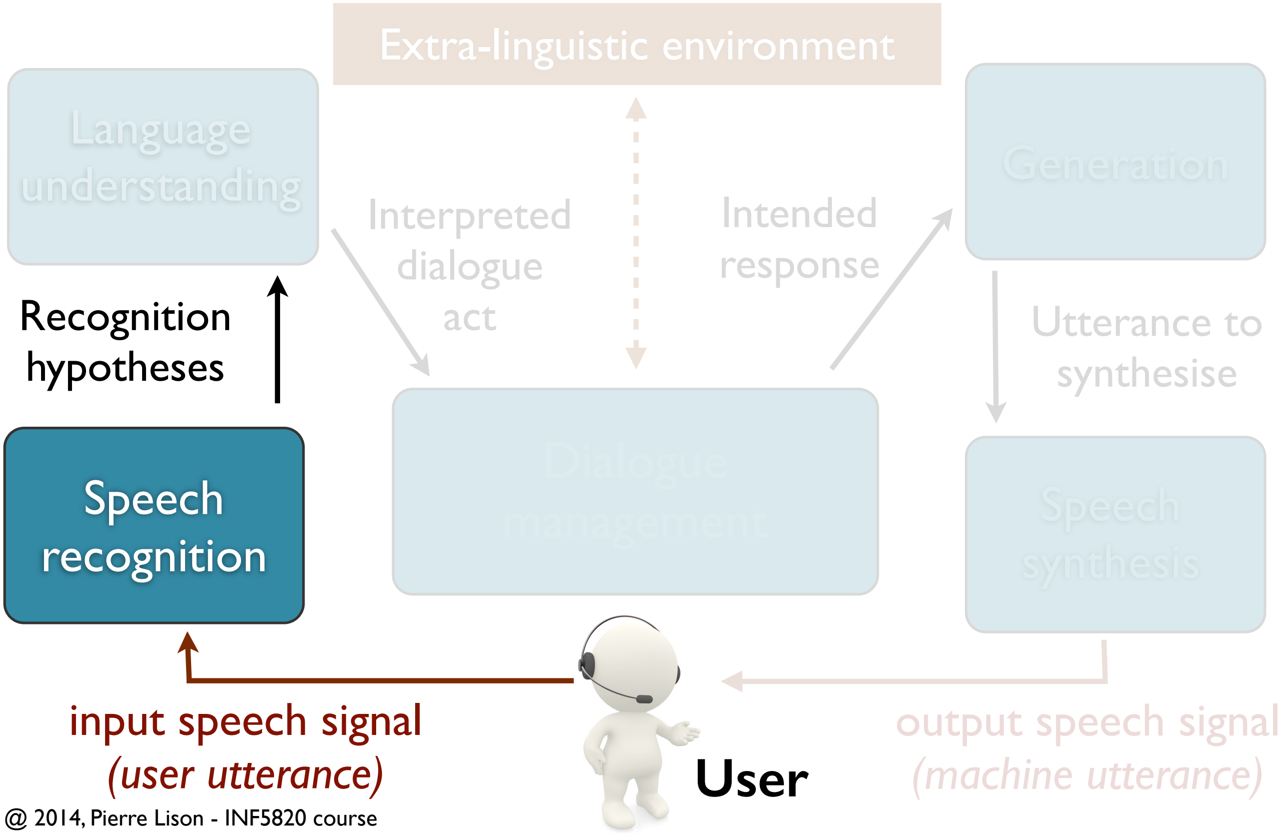
UiO : University of Oslo

INF5820:
Speech recognition

Pierre Lison,
Language Technology Group (LTG)
Department of Informatics

Fall 2014

Speech recognition



Extra-linguistic environment

Language understanding

Interpreted dialogue act

Intended response

Generation

Recognition hypotheses

Utterance to synthesise

Speech recognition

Dialogue management

Speech synthesis

input speech signal
(user utterance)

User

output speech signal
(machine utterance)

A difficult problem!

Outline

- Introduction to phonetics
- Speech recognition
- Summary

Outline

- Introduction to phonetics
  - Articulatory phonetics
  - Pronunciation variation
  - Acoustic phonetics
- Speech recognition
- Summary

Phonetics

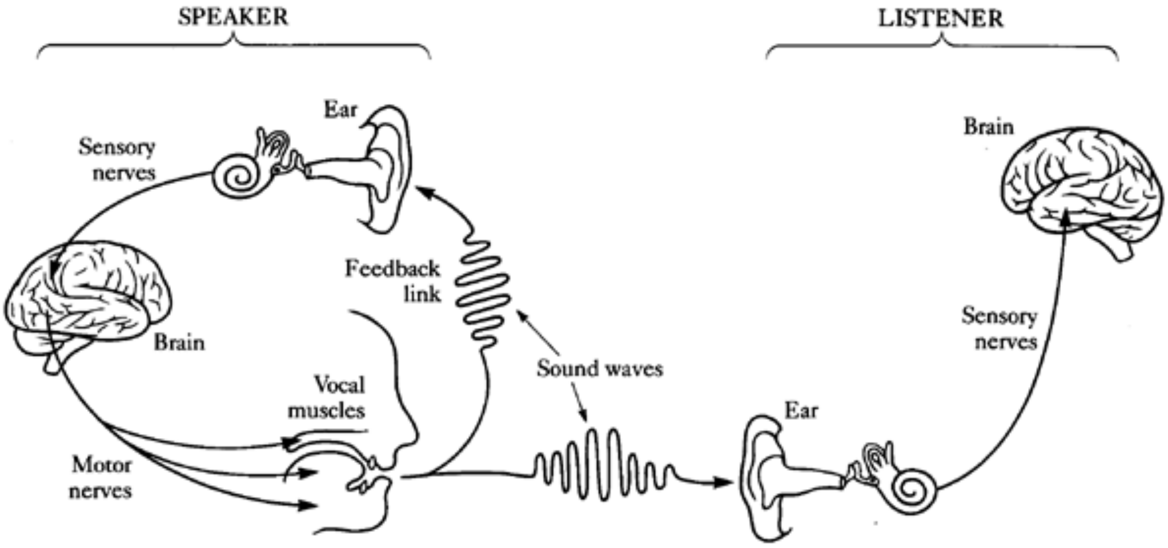
- **Phonetics** = scientific study of speech sounds
  - Divided in *articulatory*, *acoustic* and *auditory* phonetics
- The basic phonetic unit is the **phone**, which is a distinctive speech sound
- The IPA (*International Phonetic Alphabet)* is a standard for transcribing the sounds of all human languages
  - Currently 107 letters, 52 diacritics, and four prosodic marks
  - A specific language will only use a subset of these sounds
  - Compatible with other codes such as ARPAbet (U.S. English only)

IPA transcription

Bokmål	Nordavinden og sola kranglet om hvem av dem som var den sterkeste. Da kom det en mann gående med en varm frakk på seg. De blei enige om at den som først kunne få mannen til å ta av seg frakken skulle gjelde for sterkere enn den andre. Så blåste nordavinden av all si makt, men dess mer han blåste, dess tettere trakk mannen frakken rundt seg, og til sist gav nordavinden opp. Da skinte sola fram så godt og varmt, og straks tok mannen av seg frakken. Og så måtte nordavinden innrømme at sola var den sterkeste av dem.
IPA (Oslo)	[ˈˈnuːrɑˌʋinˑn̩ ɔ ˈsuːln̩ ˈˈkrɑŋlət ɔm ˈʋem ɑ dem sm̩ ˈʋɑː ɖɳ ˈˈstæɽ̊kəstə ˌdɑˑ ˈkʰɔmː de n ˈmɑnː ˌgɔˑənə me n ˈʋɑrm ˈfrɑkː pɔ ˌsæ di ble ˈˈeːnjə ɔm ɑt ˈdenː sm̩ ˈføʂʈ̠ ˌkʰʉnˑə fɔ ˈmɑnːn̩ tɔ ˈˈtʰɑː ɑ sæ ˈfrɑkːən ˌskʉlˑə ˈˈjelːə fɔ ɖɳ ˈˈstærkəstə ɑ ˌdemˑ ˈsoː ˈˈbloːstə ˈˈnuːrɑˌʋinːn̩ ɑ ˈʔɑlː sin ˈmɑkʰtʰ men ju ˈmeːɾ hɑm ˈˈbloːstə ju ˈˈtʰetːərə ˌtrɑkˑ ˈmɑnːn̩ ˈfrɑkːən ˈɾʉnt sæ ɔ tə ˈsist ˌmɔtˑə ˈˈnuːrɑˌʋinˑn̩ ˈˈjiː ˌɔpʰˑ ˌdɑː ˈˈʂintə ˈsuːln̩ ˈfrem ˌsɔˑ ˈgɔtʰː ɔ ˈʋɑrmtʰ ɑt̚ ˈmɑnːən ˈstrɑks ˌmɔtˑə ˈˈtɑː ɑ sæ ˈfrɑkːən ɔ ˈsoː ˌmɔtˑə ˈˈnuːrɑˌʋinˑn̩ ˈinːˌrømˑə ɑt ˈsuːln̩ ˈʋɑːɾ n̩ ˈˈstæɽ̊kəstə ʔɑ ˈdemː]
IPA (Fyresdal)	[ˈˈnuːrɑˌʋinˑn̩ ɔ ˈsuːlɑ ˈˈkrɑŋlɑ um ˈkʋenː ɑʊ ˌd̠æi sɔɱ ˈʋɑː d̠ən ˈˈstæɽ̊kastə ˈdoː ˈkɔmː də ɛn ˈmɑnː ˈˈgɑŋgandə mə æn ˈʋɑrmə ˈfrɑkːə pɔ ˌseˑg dæi blæɪ ˈˈʔeːnigə um ɑt dæn sɔɱ ˈfysː ˌkʰunˑə fɔ ˈmɑnːən tə ɔ ˈˈtʰɑːk ɑʊ sə ˈfrɑçːən ˌskʉlːə ˈˈɾeknɑs feɾ ɛ̃n ˈˈstærkast ɑʊ ˌd̠ɛɪ ˈsoː ˈbluːs ˈˈnuːrɑˌʋinˑn̩ ɑʊ ˈɑlː si ˈmɑkʰtʰ mən tɪ ˈmæi hɑn ˈbluːs tɪ ˈˈtʰetːɑrə ˈdruːg ˈmɑnːən ˈfrɑçːən ˈɾunt seɰ ɔ tɪ ˈs̠l̥ʉtː ˈˈmɔtːə ˈˈnuːrɑˌʋinˑn̩? ˌjeˑ ˈupːʰ ˈd̥oː ˈʂæin ˈsuːlɑ ˈfrɑmː sɔ ˈgøʰtː ɔ ˈʋɑrmt ɑt ˈmɑnːn̩ m̊e ˈʔæi ˈgɔŋ ˈˈmɔtːə ˈˈtʰɑːkə ˈɑːʊ sə ˈfrɑçːən ɔ ˈsoː ˈˈmɔtːə ˈˈnuːrɑˌʋinˑn̩ ˈʋeːˌçɛn̠ːə ʔɑt ˈsuːlɑ ˈʋɑː dən ˈˈstæɽ̊kɑ̰st ɑʊ ˈðæ̰ɪ̥]

[Nordavinden og sola, en norsk dialektprøvedatabase på nettet, NTNU, http://www.ling.hf.ntnu.no/nos/]

The speech chain



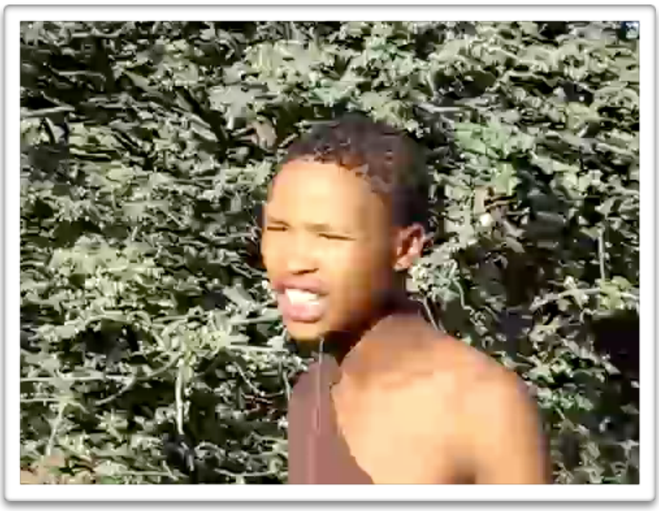
[Denes and Pinson (1993), «The speech chain»]

Speech production

- Sounds are *variations in air pressure*
- How are they produced?
  - An **air supply**: the *lungs* (we usually speak by breathing out)
  - A **sound source** setting the air in motion (e.g. vibrating) in ways relevant to speech production: the *larynx*, in which the *vocal folds* are located
  - A set of **3 filters** modulating the sound: the *pharynx*, the *oral tract* (teeth, tongue, palate,lips, etc.) & the *nasal tract*

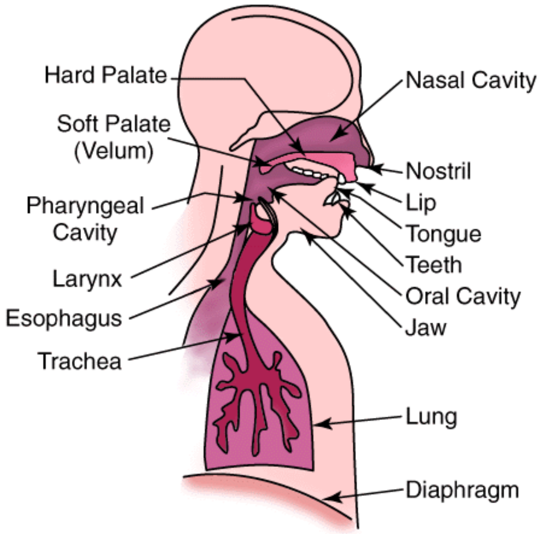
Speech production

A few languages also rely on sounds not produced by vibration of vocal folds, such as *click languages* (e.g. Khoisan family in south-east Africa):



Speech production

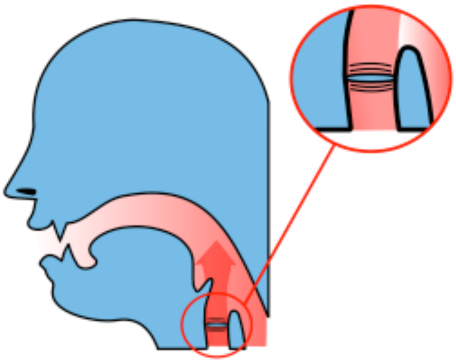
Visualisation of the vocal tract via *magnetic resonance imaging* [MRI]:



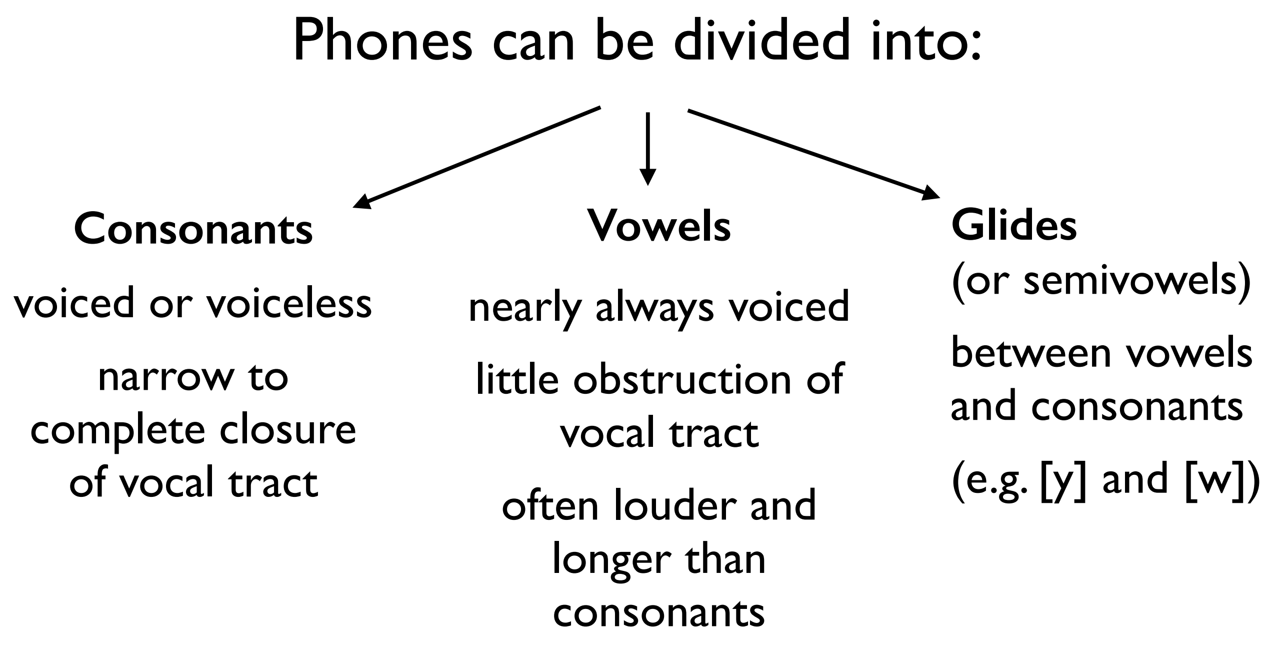
[Speech Production and Articulation Knowledge Group, University of Southern California. http://sail.usc.edu/span/]

Voiced vs. voiceless sounds

- *Voiced* sounds are made when the vocal folds are vibrating
    - e.g. [b], [d], [g], [v], [z], and all the vowels
- *Voiceless* sounds are made without such vibration
    - e.g. [p], [t], [k], [f], [s]



Sound classes



Phones can be divided into:

Consonants

voiced or voiceless

narrow to complete closure of vocal tract

Vowels

nearly always voiced

little obstruction of vocal tract

often louder and longer than consonants

Glides
(or semivowels)

between vowels and consonants

(e.g. [y] and [w])

Consonants

- Consonants are realised by restricting the airflow
- They can differ in three ways:

Voice vs voiceless
Are the vocal folds vibrating?

Manner of articulation
how does the restriction occur?

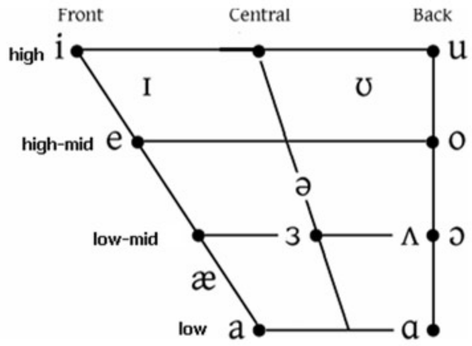
- *Stop*: airflow blocked then released: [b], [t], [k]
- *Nasal*: through nasal cavity: [n],[m],[ŋ]
- *Fricatives*: airflow constricted: [f],[v], [s]
- *Approximants*: articulators are close, but not enough to create turbulent flow: [w],[y],[l],[r]

Place of articulation
where does the restriction occur?

- *Labial* (with the lips): [p],[b],[m]
- *Coronal*: (tip or blade of tongue): [s], [t], [d]
- *Guttural* (back of the oral cavity,): [k], [g],[ŋ]

Vowels

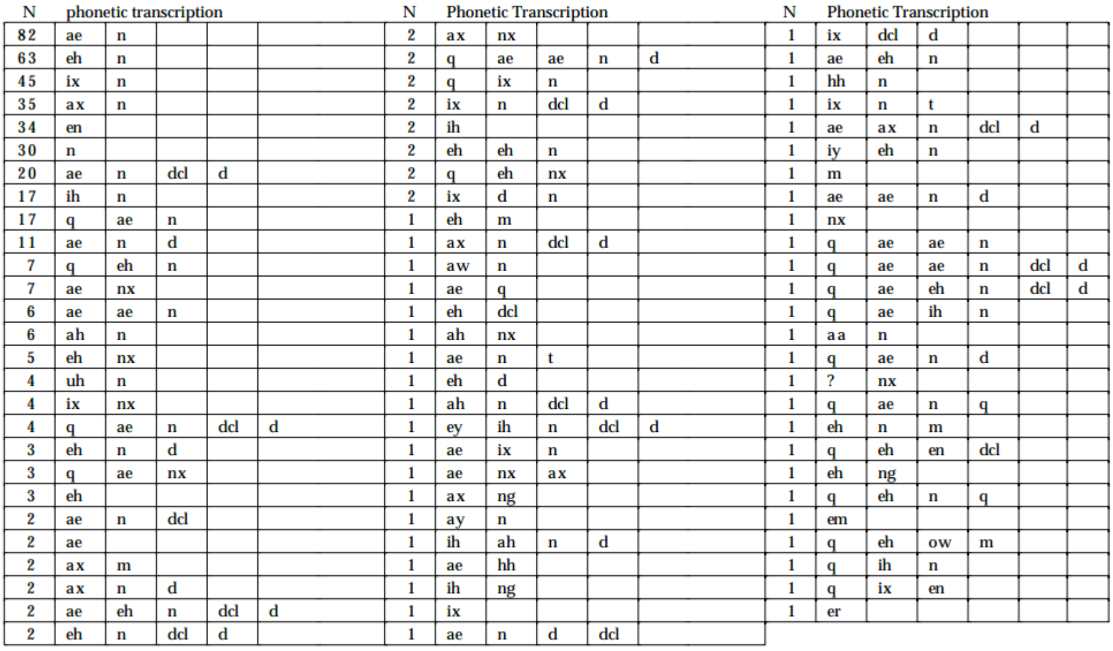
- Relevant parameters for vowels:
  - *vowel height*: height of highest part of the tongue
  - *vowel backness*: location of this high point in the oral tract
  - Shape of the lips
  - In some languages, distinction between *short* and *long* vowels



Pronunciation variation

- Words can be pronounced **very** differently:
  - *Phonetic processes*: aspirations, assimilations, deletions
  - *Co-articulation*: anticipation of next phone by articulators, or perseverance of previous move
  - Influence of various contextual factors: age, gender, environment, rate of speech, dialect, register
- Concept of **phoneme**:
  - Abstraction over a set of phones/speech sounds regarded as a single «sound» (in opposition to others) in a given language
  - Example: the English /t/ can regroup [t^h], [ɾ] and [t]

Pronunciation variation



N	phonetic transcription				
82	ae	n			
63	eh	n			
45	ix	n			
35	ax	n			
34	en				
30	n				
20	ae	n	dcl	d	
17	ih	n			
17	q	ae	n		
11	ae	n	d		
7	q	eh	n		
7	ae	nx			
6	ae	ae	n		
6	ah	n			
5	eh	nx			
4	uh	n			
4	ix	nx			
4	q	ae	n	dcl	d
3	eh	n	d		
3	q	ae	nx		
3	eh				
2	ae	n	dcl		
2	ae				
2	ax	m			
2	ax	n	d		
2	ae	eh	n	dcl	d
2	eh	n	dcl	d	

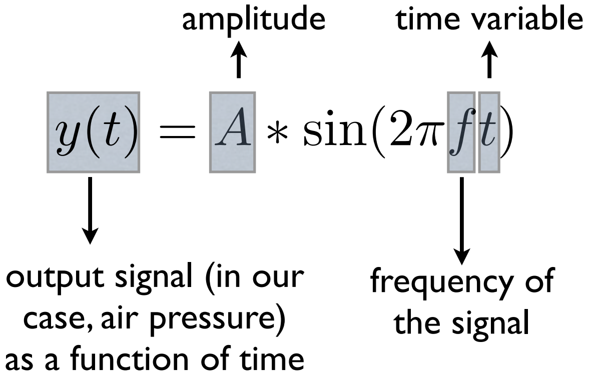
N	Phonetic Transcription				
2	ax	nx			
2	q	ae	ae	n	d
2	q	ix	n		
2	ix	n	dcl	d	
2	ih				
2	eh	eh	n		
2	q	eh	nx		
2	ix	d	n		
1	eh	m			
1	ax	n	dcl	d	
1	aw	n			
1	ae	q			
1	eh	dcl			
1	ah	nx			
1	ae	n	t		
1	eh	d			
1	ah	n	dcl	d	
1	ey	ih	n	dcl	d
1	ae	ix	n		
1	ae	nx	ax		
1	ax	ng			
1	ay	n			
1	ih	ah	n	d	
1	ae	hh			
1	ih	ng			
1	ix				
1	ae	n	d	dcl	

N	Phonetic Transcription					
1	ix	dcl	d			
1	ae	eh	n			
1	hh	n				
1	ix	n	t			
1	ae	ax	n	dcl	d	
1	iy	eh	n			
1	m					
1	ae	ae	n	d		
1	nx					
1	q	ae	ae	n		
1	q	ae	ae	n	dcl	d
1	q	ae	eh	n	dcl	d
1	q	ae	ih	n		
1	aa	n				
1	q	ae	n	d		
1	?	nx				
1	q	ae	n	q		
1	eh	n	m			
1	q	eh	en	dcl		
1	eh	ng				
1	q	eh	n	q		
1	em					
1	q	eh	ow	m		
1	q	ih	n			
1	q	ix	en			
1	er					

Table 1. 80 pronunciation variants of the word "and" from the Switchboard Transcription Corpus. The variants are listed in order of their frequency. The phonetic symbols are from a transcription system based on Arpabet. The segment [q] denotes a glottal stop. The symbol set and transcription methods are described in [15].

Acoustic phonetics

- A (speech) sound is a variation of air pressure
  - This variation originates from the speaker's speech organs
  - We can plot a *wave* showing the changes in air pressure over time (zero value being the normal air pressure)



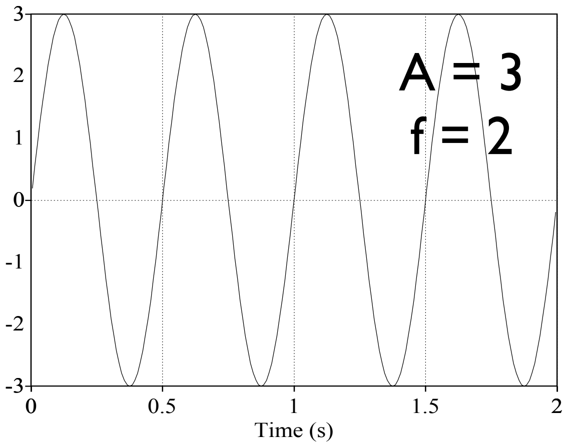
amplitude

time variable

$$\boxed{y(t)} = \boxed{A} * \sin(2\pi \boxed{f}\boxed{t})$$

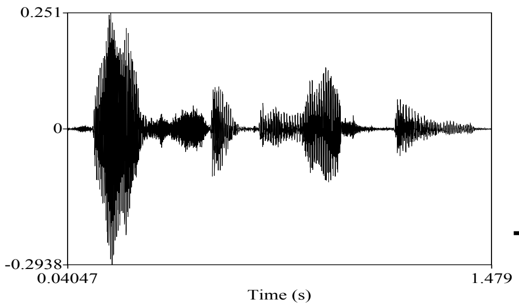
output signal (in our case, air pressure) as a function of time

frequency of the signal

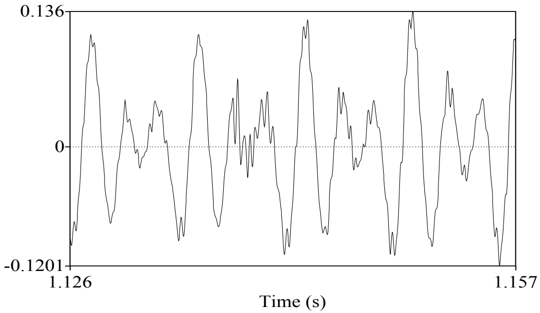


Speech waveforms

- Of course, speech is more complex than a simple sine function
- But it can still be described using the same mathematical apparatus, in terms of frequency, amplitude etc.



zoom on the part between 1.126 and 1.157 s.



can see about 4 cycles in the waveform, which means a frequency of about 4/0.03 ≈129 Hz

Signal measurements

1. The **fundamental frequency** F_0: lowest frequency of the sound wave, corresponding to the speed of vibration of the vocal folds (between 85-180 Hz for male voices and 165-255 Hz for female voices)

2. The **intensity**: the signal power normalised to the human auditory threshold, measured in **dB** (decibels):

$$\text{Intensity} = 10 \ \log_{10} \frac{\text{Power}}{P_0} = 10 \ \log_{10} \frac{1}{N P_0} \sum_{i=1}^{N} y(t_i)^2$$

for a sample of N time points $t_1, \ldots t_N$
P_0 is the human auditory threshold, = 2 x 10^{-5} Pa

Note: dB scale is logarithmic, not linear!

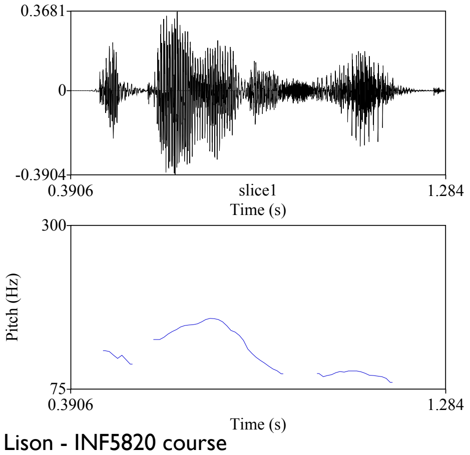
Signal measurements

Why are F_0 and the intensity important?

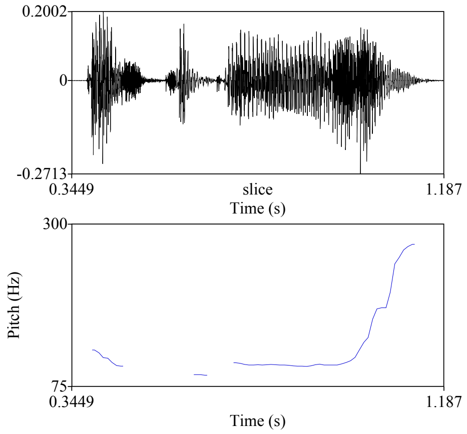
F_0 correlates with the *pitch* of the voice, and the pitch movement for an utterance will give us its *intonation*

Interrogative utterance
= rising intonation at the end

"The ball is red"



"Is the ball red?"



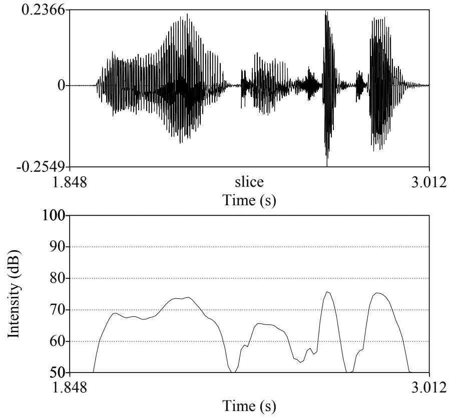
Signal measurements

Why are F_0 and the intensity important?

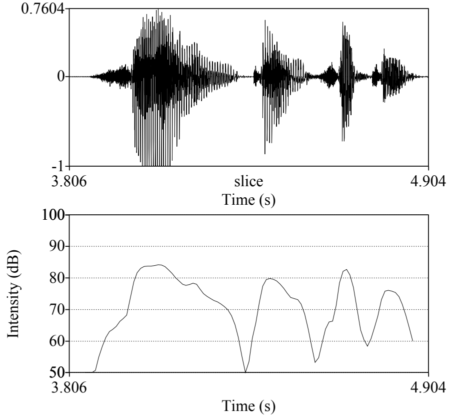
F_0
voice, and the pitch movement for an utterance will give us its *intonation*

The signal intensity corresponds to the *loudness* of the speech sound

low intensity

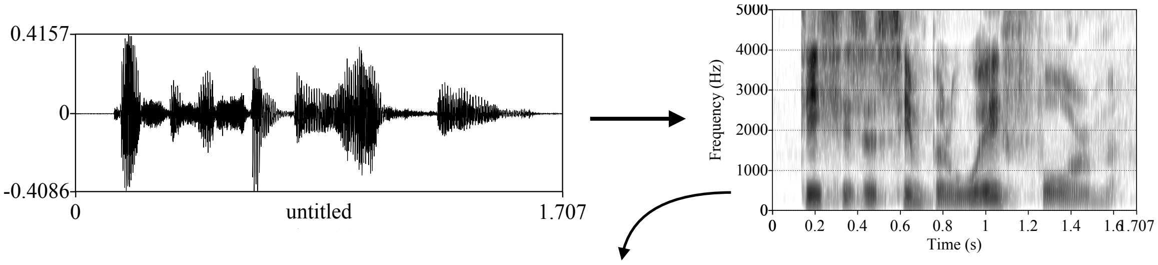


high intensity



Spectral analysis

- Possible to derive basic phonetic features (such as pitch or loudness) directly from the waveform
- But usually, the phones cannot be recognised so easily
- For this, we need to use a different representation, in terms of the signal's *component frequencies* (*spectral analysis*, based on Fourier's transform)

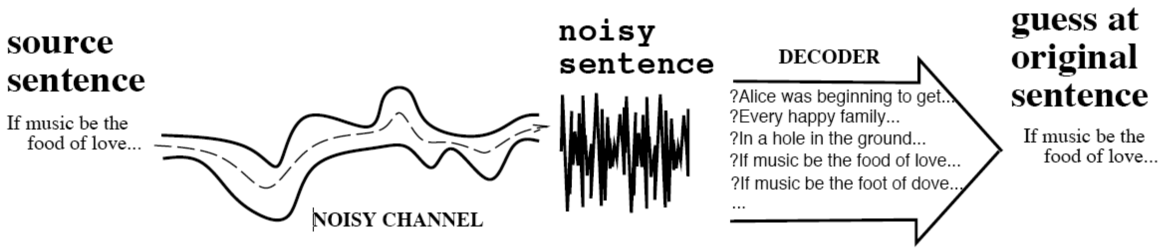


spectrogram showing how the different frequencies making up a waveform change over time

Outline

- Introduction to phonetics
- Speech recognition
  - The speech recognition problem
  - Acoustic features
  - Acoustic modelling
  - Language modelling
  - Decoding
  - Evaluation
- Summary

The noisy channel model



source sentence

If music be the food of love...

NOISY CHANNEL

noisy sentence

DECODER

?Alice was beginning to get...
?Every happy family...
?In a hole in the ground...
?If music be the food of love...
?If music be the foot of dove...
...

guess at original sentence

If music be the food of love...

- Give the observation of the noisy sentence (acoustic input), we try to guess the original sentence
- In other words, we observe the input O, and search for the best estimate W of the sentence

Formalisation

- Speech recognition as a *Hidden Markov Model*:
  - The observations is represented as a sequence of individual acoustic observations (e.g. every 10 milliseconds):

$$O = o_1, o_2, o_3, ..., o_m$$

  - The (hidden) utterance is a sequence of words:

$$W = w_1, w_2, w_3, ..., w_n$$

  - Goal: find the utterance

$$\hat{W} = \underset{W}{\operatorname{argmax}} P(W|O)$$

  - But *P(W|O)* is difficult to estimate directly!

Formalisation

- Using Bayes' rule, we can rewrite Ŵ as:

$$\hat{W} = \underset{W}{\operatorname{argmax}} \frac{P(O|W)P(W)}{P(O)} \qquad \text{(Bayes)}$$

$$= \underset{W}{\operatorname{argmax}} P(O|W)P(W) \qquad \text{(P(O) constant for all W)}$$

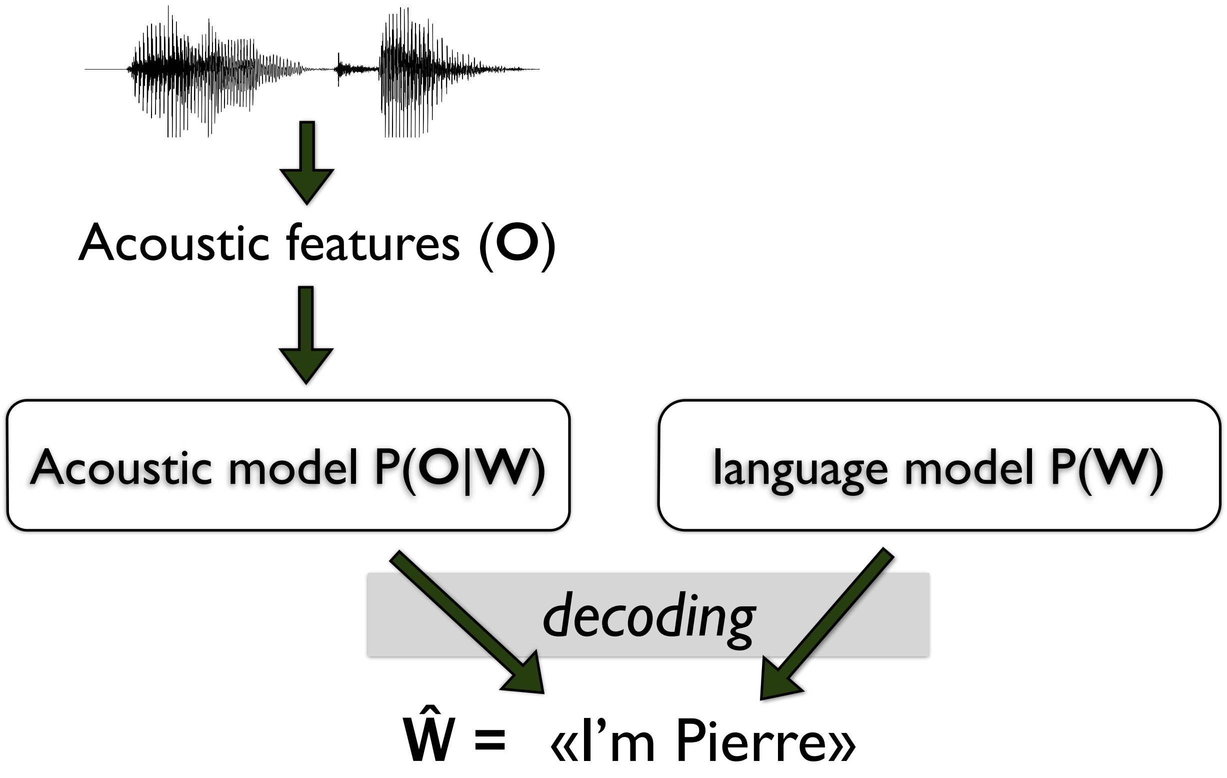
Acoustic model

Determines the probability of the acoustic inputs O given the word sequence W

Language model

Determines the probability of the word sequence W

Formalisation



Acoustic features (O)

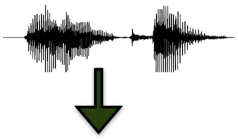
Acoustic model P(O|W)

language model P(W)

decoding

Ŵ = «I'm Pierre»

Acoustic features



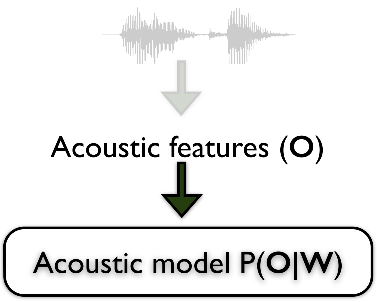
Acoustic features (**O**)

- First step: process the raw speech signal to extract its *acoustic features O*
  - The extraction is repeated at regular intervals (e.g. 10 ms)
  - The features should measure *core properties* of the signal
- *Mel-Frequency Cepstral Coefficients* (MFCC) are a popular way to extract such features
  - Total of 39 real-valued MFCC features for each time frame

MFCC steps

- **Step 1** - *Analog-to-digital conversion*: speech signal is transformed to digital form by sampling it at a given frequency (ex: 44 kHz)
- **Step 2** - *Pre-emphasis:*The amount of energy present in the high frequencies (important for speech) are boosted
- **Step 3** - *Windowing*: the signal is divided into *frames* of a given size (e.g. 10 ms). The frames might overlap (to ensure no information is lost)
- **Step 4** - *Discrete Fourier transform*: spectral analysis of the signal for each time frame (decomposition into component frequencies)
- **Step 5** - *Mel-scale wrapping*: map the DFT frequency output onto the so-called *Mel scale* (scale that model the perceptual sensitivity of the human ear)
- **Step 6** - *Cepstral analysis:* taking the log of the frequencies, and then calculating the so-called *Inverse Discrete Fourier Transform*
- **Step 7**: Derivation of all features extracted from the signal

Acoustic modelling



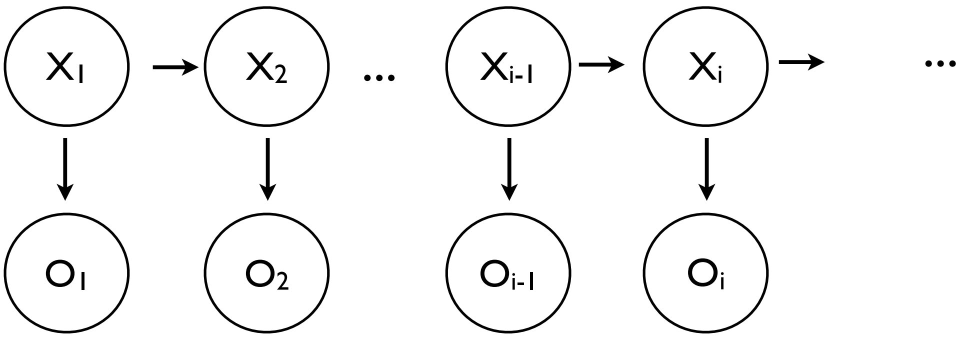
Acoustic features (O)

Acoustic model P(O|W)

- Acoustic modelling = estimation of P(O|W)
  - W represents an utterance hypothesis (word sequence)
  - O represents a sequence of acoustic features
  - The acoustic model is a *probabilistic mapping* between the acoustic features and the phones making up the utterance

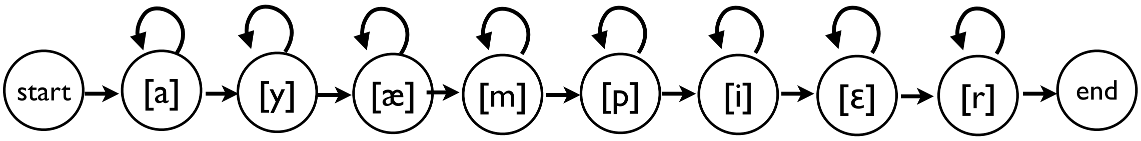
Acoustic modelling: states

- How do we estimate this distribution P(O|W)?
  - Recall that the ASR problem is essentially a Hidden Markov Model (HMM): the real utterance is «hidden» and need to be inferred from the set of acoustic observations
  - But what are exactly the *states* of the model to use?



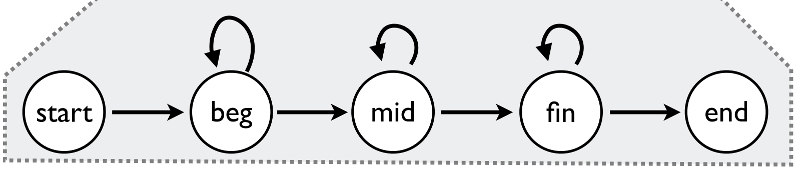
Acoustic modelling: states

- Each word can be factored into its component **phones**, corresponding to the hidden states



The arrows are the transition probabilities (observations not shown)

- Even better: define the states at the *sub-phone* level (since the spectral characteristics of a phone can vary dramatically during its pronunciation)



Often structured in three parts: **beginning, middle** and **final**

Acoustic modelling: estimation

- We are thus trying to estimate P(o|s)
  - s is a sub-phone state, e.g. the last part of the phone [b]
  - o is in the MFCC case a list of 39 real-valued features
- This distribution can be estimated from speech data, but there are two «challenges»:
  1. We don't have direct access to the exact *state* S in our data (the states are hidden)
  2. The acoustic features are real values, not discrete symbols (=infinite number of possible observations!)

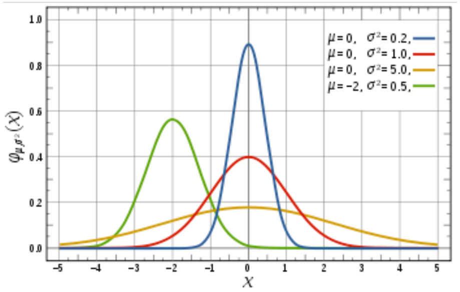
Acoustic modelling: estimation

- **First challenge**: the states are hidden!
    - *HMM training problem*: Given the observation sequence O = $o_1,\ldots o_n$, estimate both the transition probabilities between states $P(s_i|s_{i-1})$, and the observation probabilities $P(o_i|s_i)$
- It turns out that there exists a standard algorithm for estimating these probabilities
    - Called *Forward-Backward*, or *Baum-Welch* algorithm
    - Special case of a generic, iterative method called *Expectation-Maximization*

We are not going to review the details of the algorithm but see Jurafsky & Martin section 6.5 for details

Acoustic modelling: estimation

- **Second challenge**: the observations are continuous (39 real values for MFCC)
    - Acoustic models often encoded as *normal* distributions (Gaussians)
    - Each Gaussian is defined by its mean μ and variance σ^2 (both of which can be easily estimated from data)
    - To improve estimates, we can use *weighted combinations* of multiple Gaussians (= *Gaussian Mixture Models*)



Question: Assume an acoustic model with **40** different phones, MFCC features and **3** components in the Gaussian mixtures. How many parameters do we need to estimate?

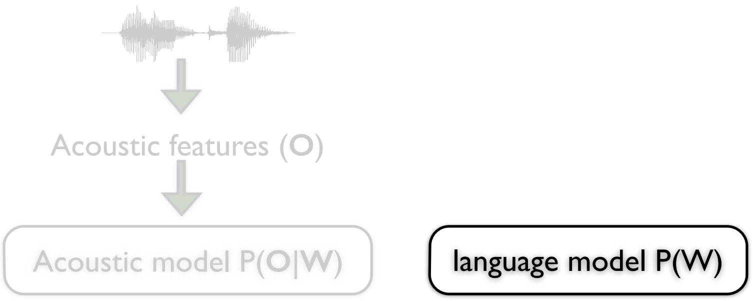
Answer: 40 *(nb. phones)* **× 39** *(nb features)* **× 3** (*nb. Gaussians*) **× 3** (*mean, variance & weight of each Gaussian*) **=14040**

Acoustic modelling: adaptation

Acoustic modelling: adaptation

- Often a mismatch between the data on which the acoustic model was trained and real-life conditions:
    - Variations in voice, accents, genre, speech rate, environmental noise, type of microphone, etc.
- But full retraining of the acoustic model is usually not feasible
- One can perform *speaker/domain adaptation* instead:
    - Generic acoustic model «wrapped» in a context-dependent model
    - Lead to huge improvements in ASR accuracy in recent years

Language modelling



Acoustic features (O)

Acoustic model P(O|W)

language model P(W)

- The second ASR model is the *language model*
  - Most interesting part for us: we rarely touch the system's acoustic models, but need to provide the domain's language model(s)
- Encodes the likelihood of an utterance P(W)
  - Mapping from words to possible pronunciations is done using a phonetic dictionary: sterkeste -> ˈˈstær̥kəstə

Language modelling

- First option: define P(W) with a hand-crafted *grammar*
  - The (context-free) grammar must specify all possible utterances for the domain

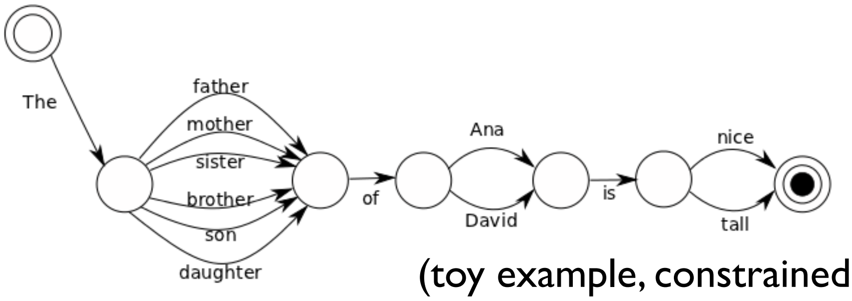
+	Good accuracy
-	Highly rigid

- Alternatively, one could estimate a *statistical model* from data
  - Need to collect data from the interaction (via e.g. Wizard-of-Oz interactions)

+	More flexible
-	Higher WER

Recognition grammars

- Explicit specification of possible utterances
  - Usually some variant of context-free grammars
  - Can include weights to increase/decrease the likelihood of particular phrases
  - Everything not covered by the grammar is ignored!



(toy example, constrained here to a finite-state)

Recognition grammars

```
<grammar version="1.0" root="simpleGrammar">
<!--Simple grammar for greetings and closings -->

<rule id="simpleGrammar">
    <one-of>
        <ruleref uri="#greeting"/>
        <ruleref uri="#closing"/>
    </one-of>
</rule>

<rule id="greeting">
    <item>
        <one-of>
            <item>hello</item>
             <item>hi</item>
      </one-of>
    </item>
   <item repeat="repeat="0-1">
        <ruleref uri= "#name "/>
    </item>
</rule>

      ....
```

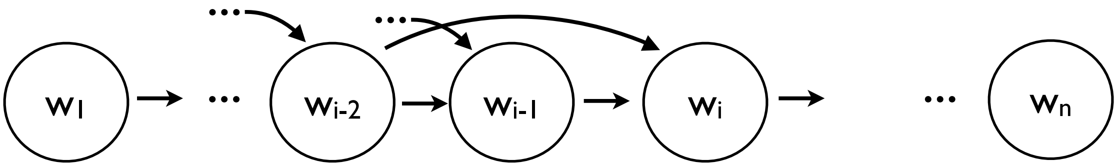
```
 ....

<rule id="closing">
    <item>
        <one-of>
            <item>goodbye</item>
             <item>bye</item>
      </one-of>
    </item>
   <item repeat="repeat="0-1">
        <ruleref uri= "#name "/>
    </item>
</rule>

<rule id= "name ">
    <one-of>
        <item>Pierre</item>
        <item>Jan Tore</item>
    </one-of>
</rule>
</grammar>
```

Statistical language modelling

- How can we estimate P(W) from data?
- The probability P(W) is often factored as a *Markov Chain* of order k, also called an *N-gram*:

$$P(w_i|w_{i-1},..., w_1) = P(w_i| w_{i-1},... w_{i-k})$$

- Generally, k=2 or 3 (bigram or trigram)



(here for k=2)

Statistical language modelling

- The probabilities $P(w_i|w_{i-1},...w_{i-k})$ can be estimated by counting relative occurrences in the data
- However, «plain» estimation has a problem with low-frequency counts:
  - If the sequence $[w_{i-k},..w_{i-1},w_i]$ never occurs in the data, the probability $P(w_i|w_{i-1},...w_{i-k})$ will be set to zero
  - Not a reasonable assumption with limited training data!
  - **Solution**: use *smoothing techniques* (e.g. Good-Turing)

See chap. 4 of Jurafsky & Martin for details

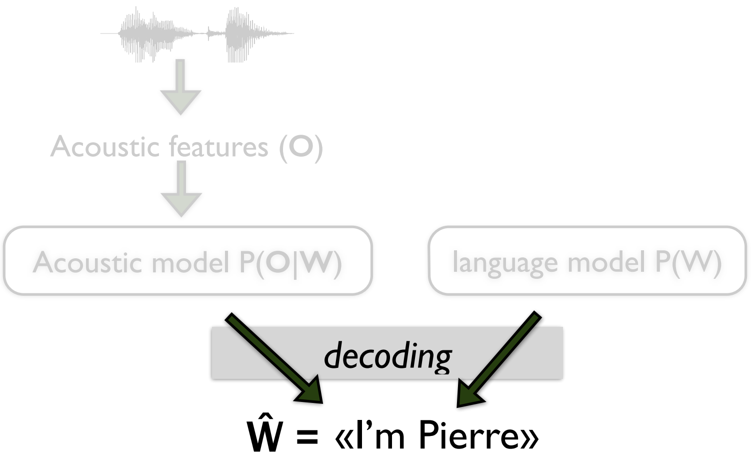
Language model adaptation

- Some speech recognisers allow the language model to be modified «on the fly», at runtime
- We can exploit the context to «prime» parts of the model, and adapt the probabilities to the situation
  - Example: if the system just asked a yes/no question, the user is more likely to answer «yes» or «no»
  - Example: if the environment contains certain objects (e.g. a box), the user is more likely to mention them
- Can lead to big improvements in accuracy

[P. Lison. A salience-driven approach to speech recognition for human-robot interaction, *Interfaces: Explorations in Logic, Language and Computation*, 2010]

Decoding

- Decoding = *search* for the most likely word sequence given the observations
- Large search space, but there are various «tricks» to speed up the search (e.g. *dynamic programming*)



Decoding

- Specialised algorithms exist to perform this kind of search operations efficiently
- Most well-known decoding algorithm is *Viterbi*
  - Viterbi processes the observation sequence left to right and calculates the state probabilities at the current step based on the previous step
  - Other algorithms also possible to perform e.g. multi-pass decoding (generate N-best results)

See section 6.4, 9.6 and 10.1 of Martin & Jurafsky for details

ASR evaluation

- Standard evaluation metric: *Word Error Rate*
  - Measures how much the utterance hypothesis *h* differs from the «gold standard» transcription t^*
- Relies on a minimum edit distance between *h* and t^*, counting the number of word *substitutions*, *insertions* and *deletions*:

$$\text{Word Error Rate} = 100 \times \frac{\text{Insertions} + \text{Substitutions} + \text{Deletions}}{\text{Number of words in transcription}}$$

ASR evaluation

- Examples of evaluation:

Gold standard Transcription	*yes can you now rotate this triangle*
ASR hypothesis	*yes can you not rotate this triangle there*

$$\mathrm{WER} = 100 \times \frac{\text{1 Sub} + \text{1 Ins}}{7}$$

$$= 28.6\%$$

Gold standard Transcription	*there is five and*
ASR hypothesis	*the size and*

$$\mathrm{WER} = 100 \times \frac{\text{2 Sub} + \text{1 Del}}{4}$$

$$= 75\%$$

Summary

- We introduced some basic concepts of phonetics, such as phones, pitch or loudness
- **Speech recognition**: Find ("decode") the *most likely utterance(s)* for a speech signal, based on two probabilistic models:
    - *Acoustic model*: likelihood of observed acoustic features for each possible (sub-)phone
    - *Language model*: likelihood of particular sequence of phones
- Evaluation of ASR results via edit-distance metric

Next lecture

- Next Friday, we'll move to the next step in the processing pipeline: natural language *understanding* (NLU)
- We'll review some of the core tasks that need to be achieved there:
  - Correction of disfluencies
  - Semantic parsing
  - Reference resolution
  - Dialogue act recognition