Language Technology
http://cs.lth.se/edan20/
Chapter 19: Speech Recognition

Pierre Nugues

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

October 14, 2021

Speech Recognition

Conditions to take into account:

- Number of speakers
- Fluency of speech.
- Size of vocabulary
- Syntax
- Environment

Structure of Speech Recognition

Words:

$$W = w_1, w_2, \ldots, w_n.$$

Acoustic symbols:

$$A = a_1, a_2, \ldots, a_m,$$

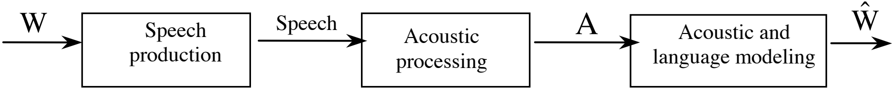
$$\hat{W} = \arg\max_{W} P(W|A).$$

Using Bayes' formula,

$$P(W|A) = \frac{P(A|W)P(W)}{P(A)}.$$

Two-Step Recognition

$$\hat{W} = \arg\max_{W} P(A|W)P(W).$$



Speech Parameters

Recognition devices derive a set of acoustic parameters from speech frames.
Parameters should be related to "natural" features of speech: voiced or unvoiced segments.
A simple parameter giving a rough estimate of it: the energy: the darker the frame, the higher the energy.

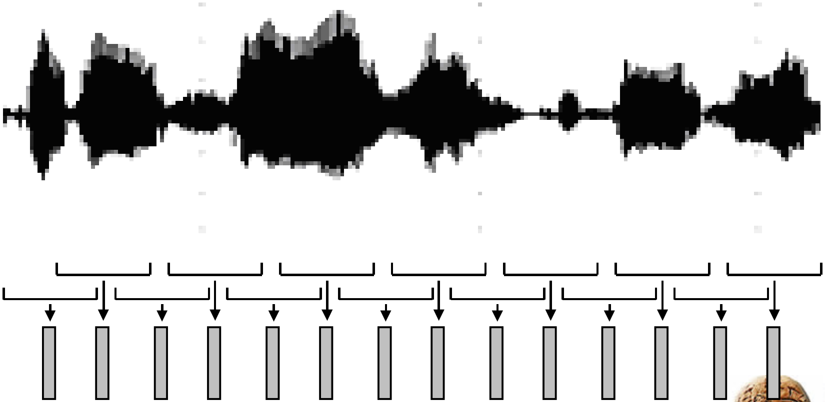
$$E(F_k) = \sum_{n=m}^{m+N-1} s^2(n).$$

Linear prediction coefficients:

$$\hat{s}(n) = a(1)s(n-1) + a(2)s(n-2) + a(3)s(n-3) + ... + a(m)s(n-m),$$

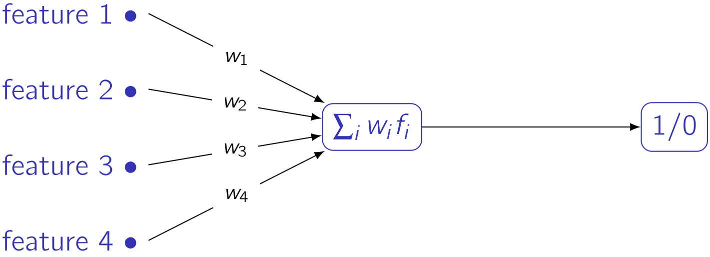
Extraction of Speech Parameters

Features are extracted every 10 ms over a 20 s frame



Neural Networks: Representation

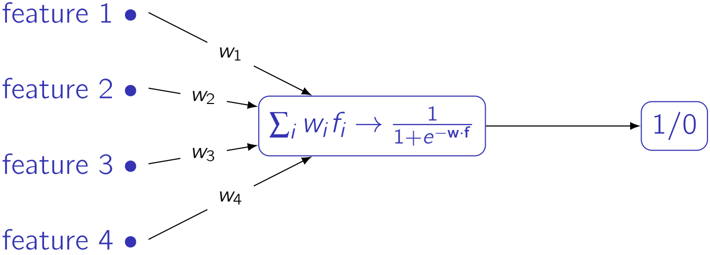
Another representation of the perceptron:



The base network: An input layer and an output layer

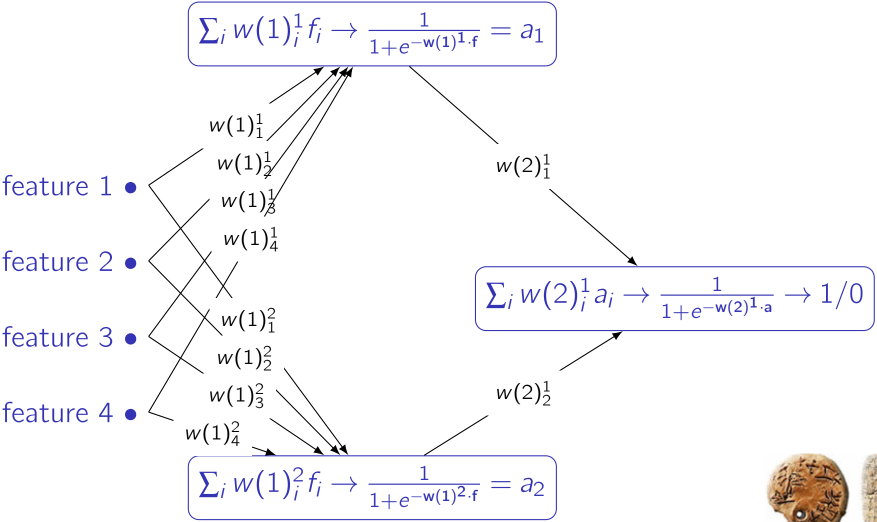
Neural Networks: Activation Function

And logistic regression:



The logistic function is the activation function of the node

Neural Networks: Hidden Layers



Word Decoding

Markov models are a probabilistic mapping of a string of acoustic symbols $a_1, a_2, ..., a_m$ onto a string of phonemes $\varphi_1, \varphi_2, ..., \varphi_m$.

A language model applies a second probability to a word sequence.

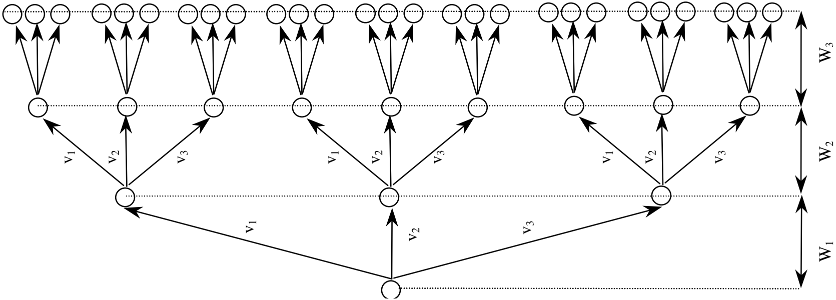
The complete speech recognition then consists in decoding word sequences $w_1, w_2, ..., w_n$ from phonemic strings and weighting them using the language model.

$$
\begin{array}{c} \textit{words} \\ \textit{phonemes} \\ \textit{ac.symbols} \end{array}
\quad
\overbrace{\begin{array}{ccc} \varphi_1^1 & \varphi_1^2 & \varphi_1^{m1} \\ a_1^1, & a_1^2, \ldots, & a_1^{m1} \end{array}}^{w_1},
\overbrace{\begin{array}{ccc} \varphi_2^1 & \varphi_2^2 & \varphi_2^{m2} \\ a_2^1, & a_2^2, \ldots, & a_2^{m2} \end{array}}^{w_2}, \ldots,
\overbrace{\begin{array}{cc} \varphi_j^1 & \varphi_j^{mj} \\ a_j^1, \ldots, & a_j^{mj} \end{array}}^{w_j}, \ldots
$$

Searching Words

A hypothesis search.

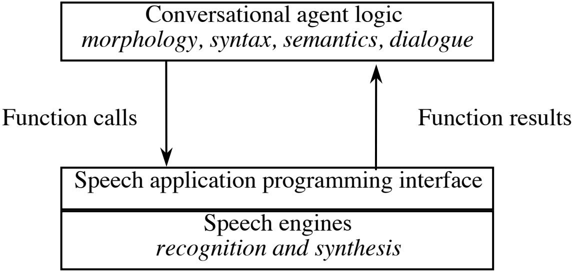
If the vocabulary contains k words $v_1, v_2, \ldots, v_k$, w_1 is to be selected amongst k possibilities, w_2 amongst k possible choices again and so on.



Decoding uses the A* algorithm

Commercial Systems

Speech recognition systems are accessible using an API



In addition to a language model, speech engines often give the possibility to use a phrase–structure grammar