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 17, 2019

Speech Recognition

Conditions to take into account:

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

Structure of Speech Recognition

Words:

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

Acoustic symbols:

$$A = a_1, a_2, ..., a_m,$$

$$\hat{W} = \underset{W}{\arg\max}\, P(W|A).$$

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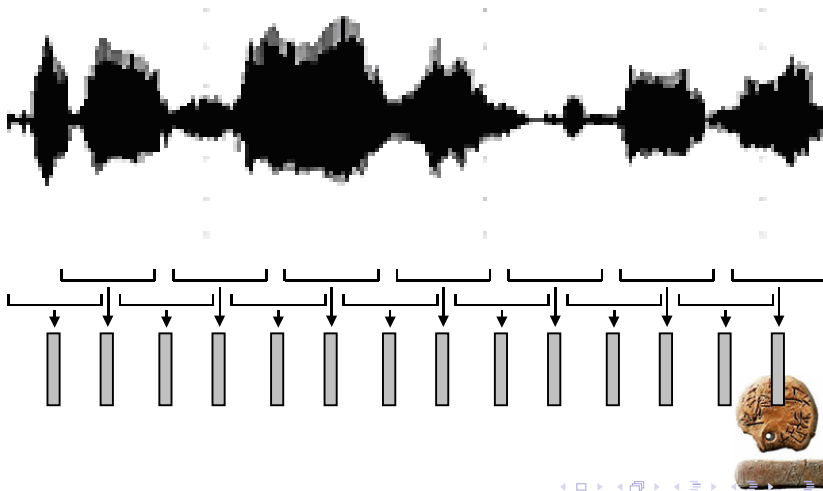
$$E(F_k) = \sum_{n=m}^{m+N-1} s^2(n).$$

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:

feature 1 •

w_1

feature 2 •

w_2

$\sum_i w_i f_i$

$1/0$

feature 3 •

w_3

w_4

feature 4 •
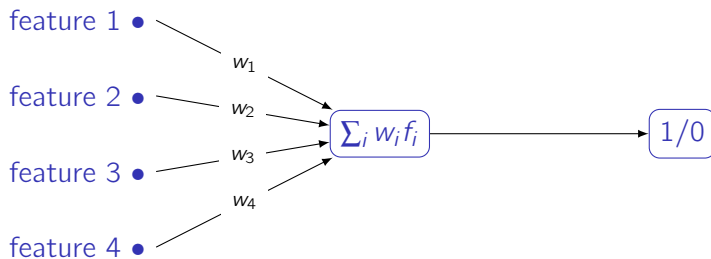
The base network: An input layer and an output layer

Neural Networks: Activation Function

And logistic regression:

feature 1 •

feature 2 • w_1

feature 3 • w_2 $\sum_i w_i f_i \rightarrow \frac{1}{1+e^{-\mathbf{w}\cdot\mathbf{f}}}$ 1/0

feature 4 • w_3

w_4
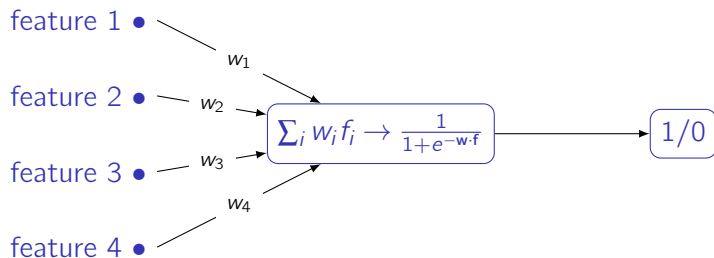
The logistic function is the activation function of the node

Neural Networks: Hidden Layers

$$\sum_i w(1)_i^1 f_i \rightarrow \frac{1}{1+e^{-\mathbf{w(1)^1 \cdot f}}} = a_1$$

$w(1)_1^1$

$w(1)_2^1$

$w(1)_3^1$

$w(1)_4^1$

$w(2)_1^1$

feature 1 •

feature 2 •

$w(1)_1^2$

feature 3 •

$w(1)_2^2$

$\sum_i w(2)_i^1 a_i \rightarrow \frac{1}{1+e^{-\mathbf{w(2)^1 \cdot a}}} \rightarrow 1/0$

$w(1)_3^2$

feature 4 •

$w(2)_2^1$

$w(1)_4^2$

$$\sum_i w(1)_i^2 f_i \rightarrow \frac{1}{1+e^{-\mathbf{w(1)^2 \cdot f}}} = a_2$$

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.

$$
\begin{array}{ccccccccccc}
\text{words} & \overbrace{\phantom{\varphi_1^1 \quad \varphi_1^2 \quad \varphi_1^{m1}}}^{w_1} & & \overbrace{\phantom{\varphi_2^1 \quad \varphi_2^2 \quad \varphi_2^{m2}}}^{w_2} & & & \overbrace{\phantom{\varphi_j^1 \quad \varphi_j^{mj}}}^{w_j} & \\
\text{phonemes} & \varphi_1^1 & \varphi_1^2 & \cdots & \varphi_1^{m1} & \varphi_2^1 & \varphi_2^2 & \cdots & \varphi_2^{m2} & \cdots, & \varphi_j^1 & \cdots, & \varphi_j^{mj} & \cdots \\
\text{ac.symbols} & a_1^1 & , & a_1^2 & , \cdots & a_1^{m1} & , & a_2^1 & , & a_2^2 & \cdots & a_2^{m2} & , \cdots, & a_j^1 & , \cdots, & a_j^{mj} & , \cdots
\end{array}
$$
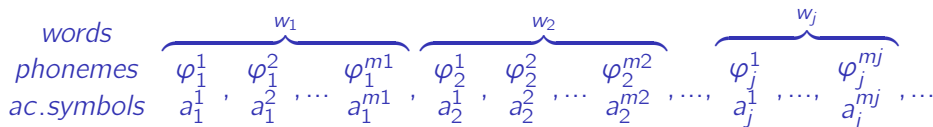
Searching Words

A hypothesis search.

If the vocabulary contains k words $v_1, v_2, ..., v_k$, 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


```
┌─────────────────────────────────────┐
│   Conversational agent logic         │
│ morphology, syntax, semantics, dialogue │
└─────────────────────────────────────┘

Function calls                    Function results

┌─────────────────────────────────────┐
│ Speech application programming interface │
├─────────────────────────────────────┤
│          Speech engines              │
│      recognition and synthesis       │
└─────────────────────────────────────┘
```

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