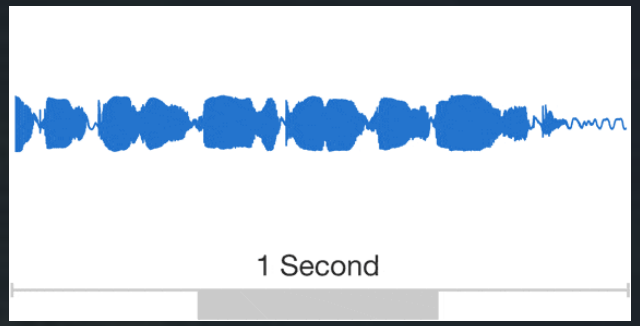
Artificial voice production

Martin Eeg, statistician, Anders Jønsson



Introduction:

- Speech synthesis: When computers generate speech

  - The term Speech synthesis refers to the technologies that enable computers or other electronic systems to output simulated human speech
  - Important are: intelligibility and naturalness
  - Naturalness is often evaluated depending on every situation

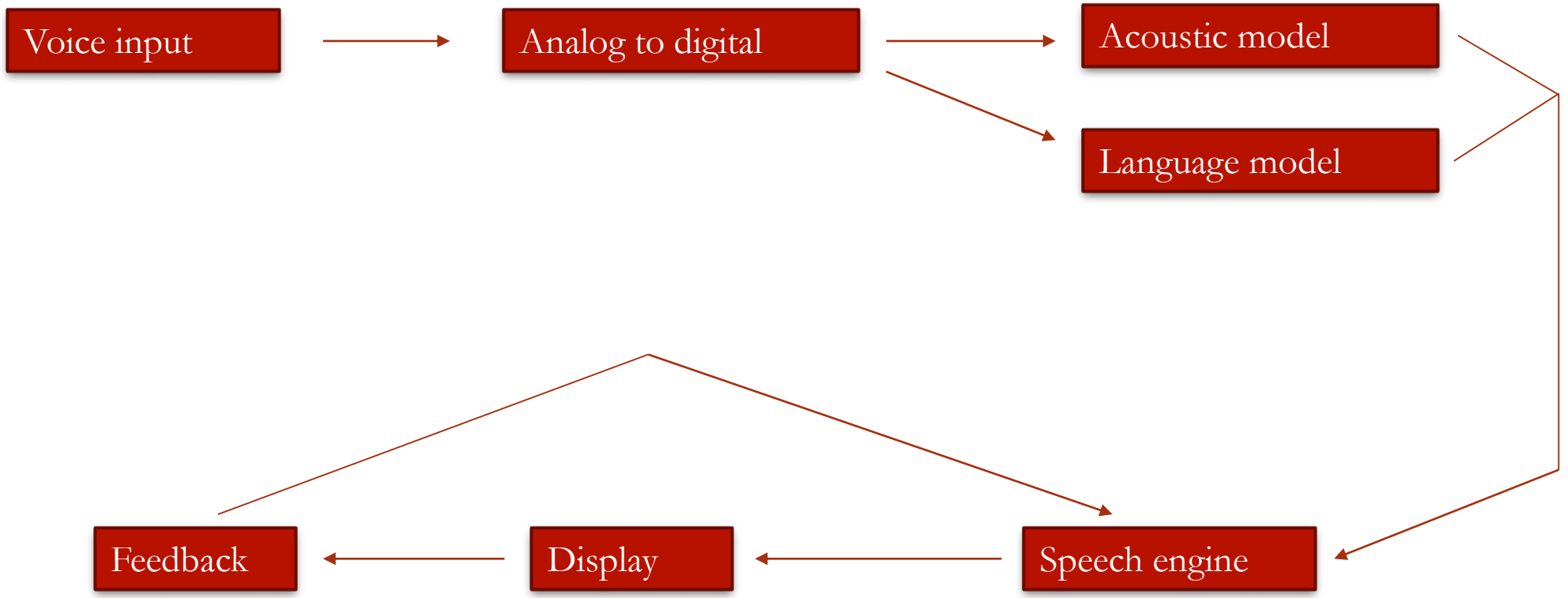
Types of Speech Synthesis:

- Concatenative
  - Based on human speech samples
    - Diphones (sound to sound transitions, German has about 2.500 and Spanish about 800)
    - Words
    - Variable length units
- Formant Synthesis
  - Simulates human speech electronically using phonological rules

TTS: Text – to – Speech

- Translates text into speech using phonetic rules to transcribe the text and then speak it.
- Requires information about:
  - Abbreviations (Dr., Nr., etc., …)
  - Specific readings of numbers and symbols ($, #, %, &, @)
  - Reading of time formats (1:45, 13;45…)
  - Pronunciation of each letter in every context (“map”, “pap”, “Jane”,…)

Speech Recognition:



Charu Joshi, 2010

Acoustic model:

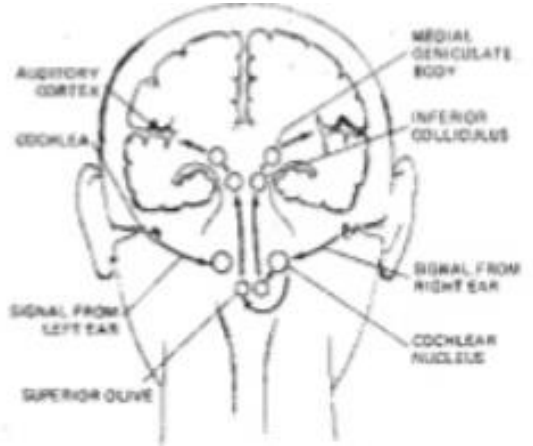
- Is created by taking audio recordings of speech, and their text transcriptions, and using software to create statistical representations of the sounds that make up each word. It is used by a speech recognition engine to recognize speech

Language model:

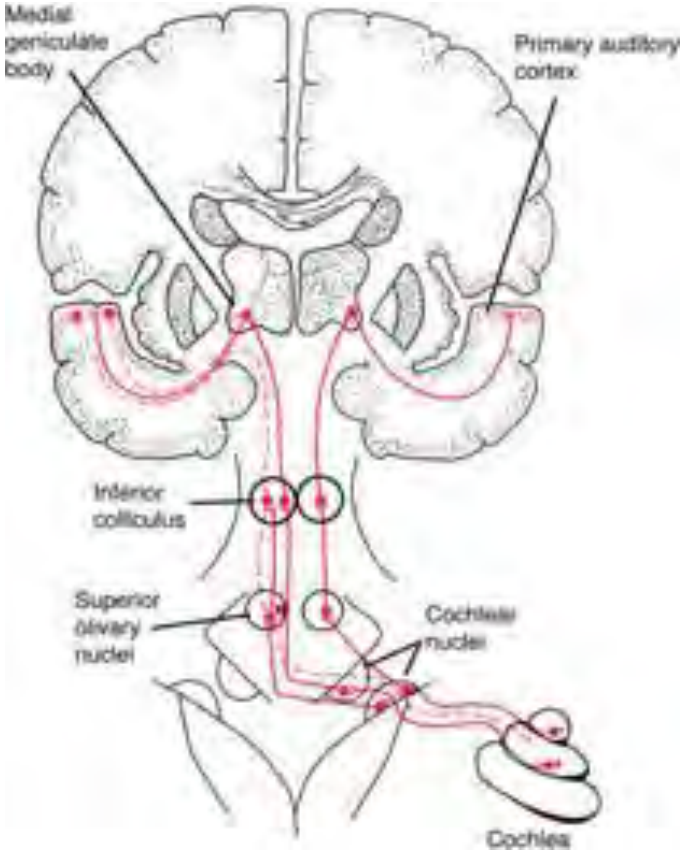
- Is used in many natural language processing applications such as speech recognition tries to capture the properties of a language, and to predict the next word in a speech sequence

How do humans do it?:

- Articulation produces sound waves which the ear conveys to the brain for processing

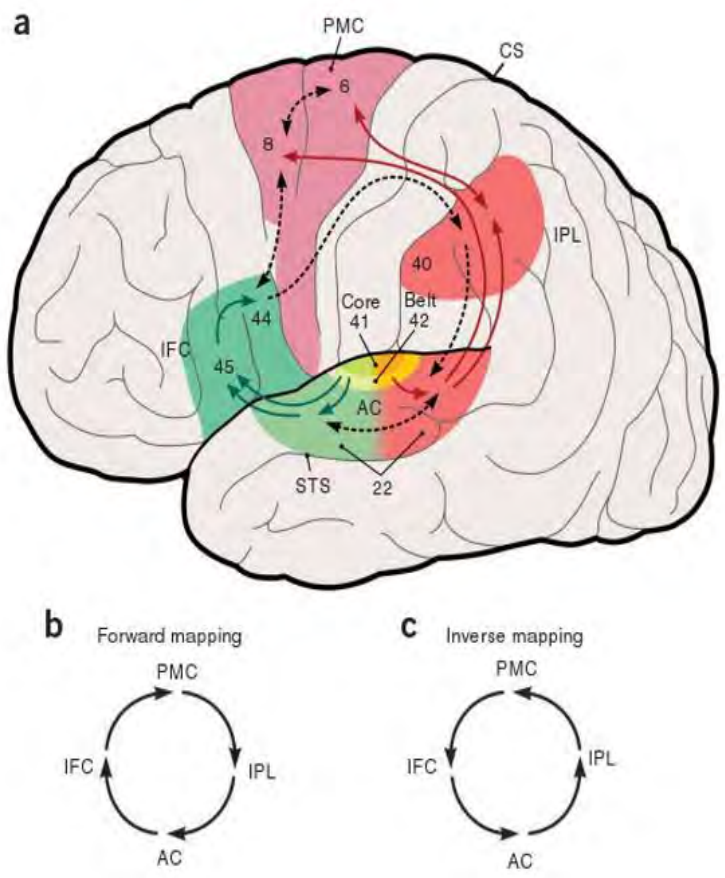


The hearing anatomy



- Nerve signals propagate with a rate of approximately 10-30 m/s and will therefore be delayed up through the hearing pitch.
  In each synapse is a further delay of approximately 1 ms.

Main brain regions involved in auditory processing



Dual auditory processing scheme of the human brain and the role of internal models in sensory systems.
This expanded scheme closes the loop between speech perception and production and proposes a common computational structure for space processing and speech control in the postero-dorsal auditory stream.
(a) Antero-ventral (green) and postero-dorsal (red) streams originating from the auditory belt. The postero-dorsal stream interfaces with premotor areas and pivots around inferior parietal cortex, where a quick sketch of sensory event information is compared with a predictive efference copy of motor plans.
(b) In one direction, the model performs a forward mapping: object information, such as speech, is decoded in the antero-ventral stream all the way to category-invariant inferior frontal cortex (area 45), and is transformed into motor-articulatory representations (area 44 and ventral **PMC**), whose activation is transmitted to the IPL (and posterior superior temporal cortex) as an **efference copy**. **(c) In reverse direction, the model performs an inverse mapping, whereby attention- or intention-related changes in the IPL influence the selection of context-dependent action programs in PFC and PMC.**
Both types of dynamic model are testable using techniques with high temporal precision (for example, magnetoencephalography in humans or single-unit studies in monkeys) that allow determination of the order of events in the respective neural systems.
AC, auditory cortex; STS, superior temporal sulcus; **IFC**, inferior frontal cortex, PMC, premotor cortex; **IPL**, inferior parietal lobule; **CS**, central sulcus. Numbers correspond to Brodmann areas.

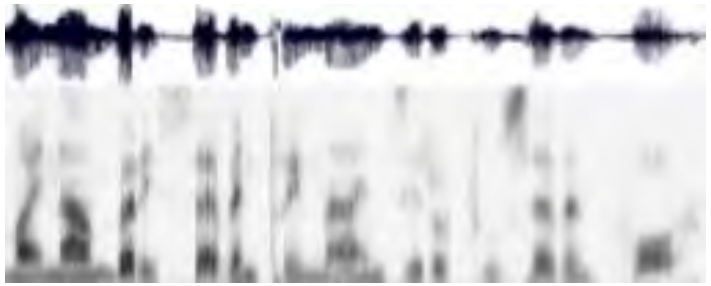
Rauschecker J, Scott S. Maps and streams in the auditory cortex: nonhuman primates illuminate human speech processing. 2009. Nat Neurosci. Vol. 12(6), pp. 718-724.

How might computers do it?:

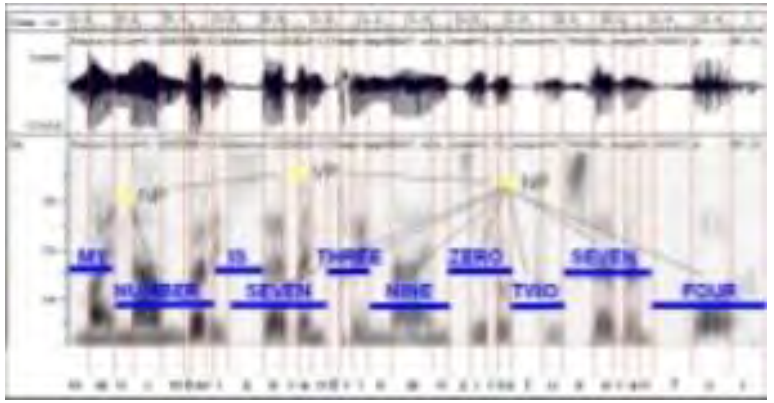
- Acoustic waveform

- Digitization
- Acoustic analysis of the speech signal
- Linguistic interpretation

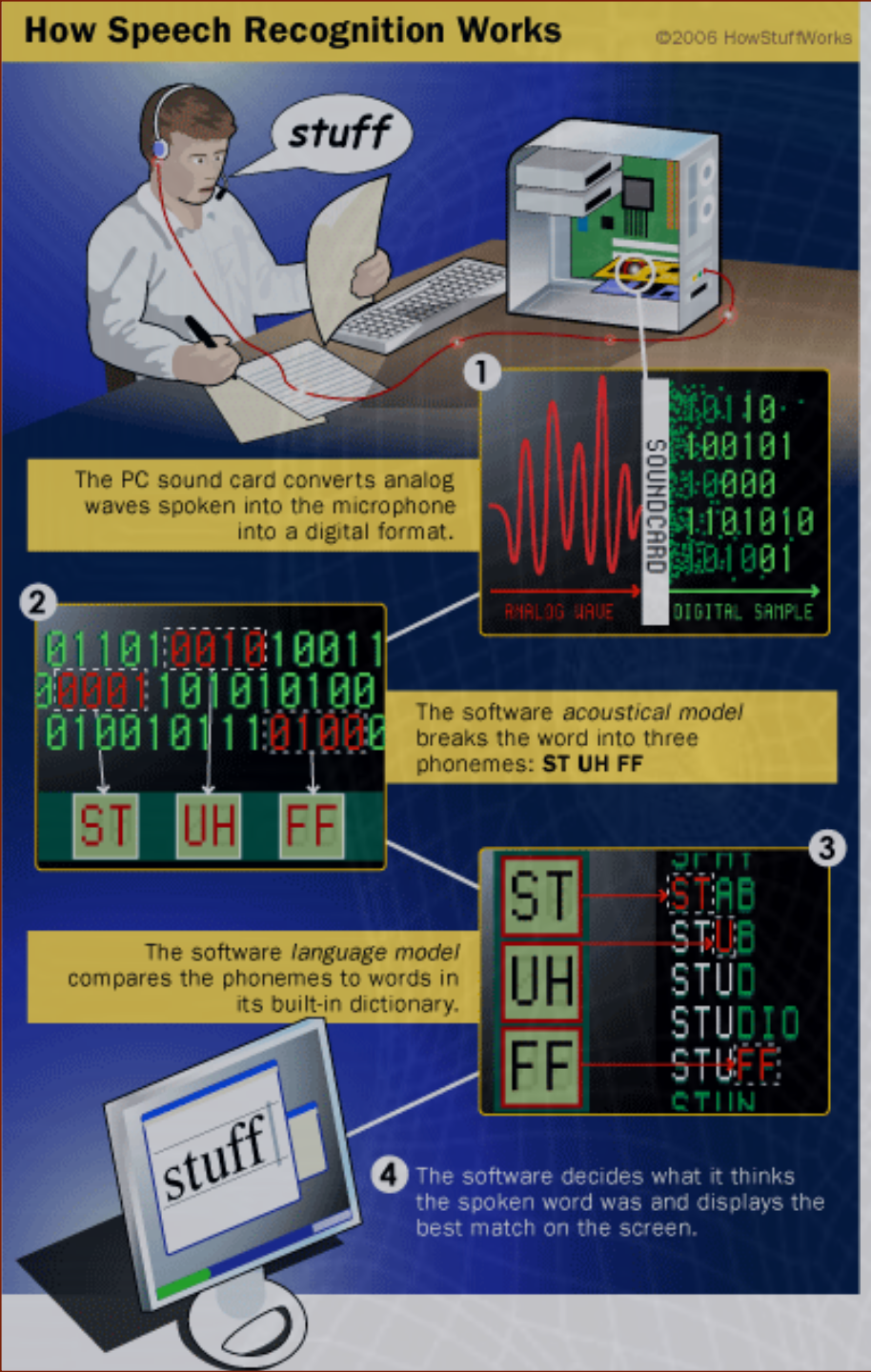
- Acoustic signal

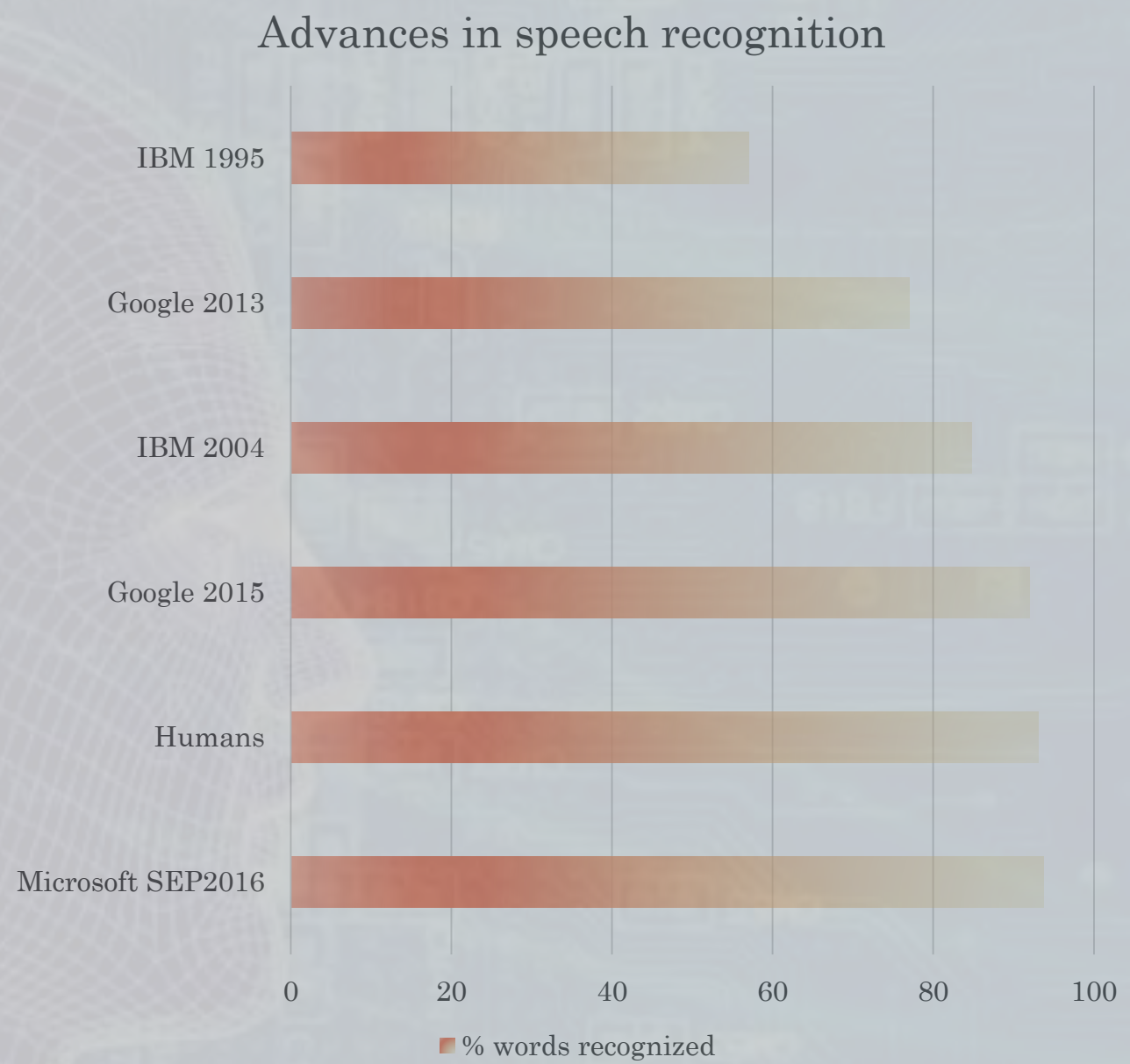


- Speech recognition



Charu Joshi, 2010


How Speech Recognition Works
©2006 HowStuffWorks
stuff
1
SOUNDCARD
ANALOG WAVE
DIGITAL SAMPLE
The PC sound card converts analog waves spoken into the microphone into a digital format.
2
ST
UH
FF
The software *acoustical model* breaks the word into three phonemes: **ST UH FF**
3
ST
UH
FF
STAB
STUB
STUD
STUDIO
STUFF
The software *language model* compares the phonemes to words in its built-in dictionary.
stuff
4
The software decides what it thinks the spoken word was and displays the best match on the screen.


Advances in speech recognition
IBM 1995
Google 2013
IBM 2004
Google 2015
Humans
Microsoft SEP2016
0
20
40
60
80
100
% words recognized


the great performance

Artificial voice production

Martin Eeg, statistician, Anders Jønsson