Automatic speech recognition

- Content today:
  - Collection of speech data
  - ASR systems today
  - ASR applications
  - ASR courses
- Presented by Mikko Kurimo

About Mikko

- Background from Machine Learning algorithms and Pattern Recognition systems
- PhD 1997 at TKK on speech recognition systems
- Postdoc research at several speech groups:
  - IDIAP, Martigny, CH
  - Univ. Colorado and Stanford Research Center, USA
  - Univ. Edinburgh, UK and Univ. Saint Etienne, FR
- Head of TKK's speech recognition and multimodal interfaces research groups, several national and European speech projects
- Large-vocabulary continuous speech recognition, language modeling, information retrieval from speech

Goals of today

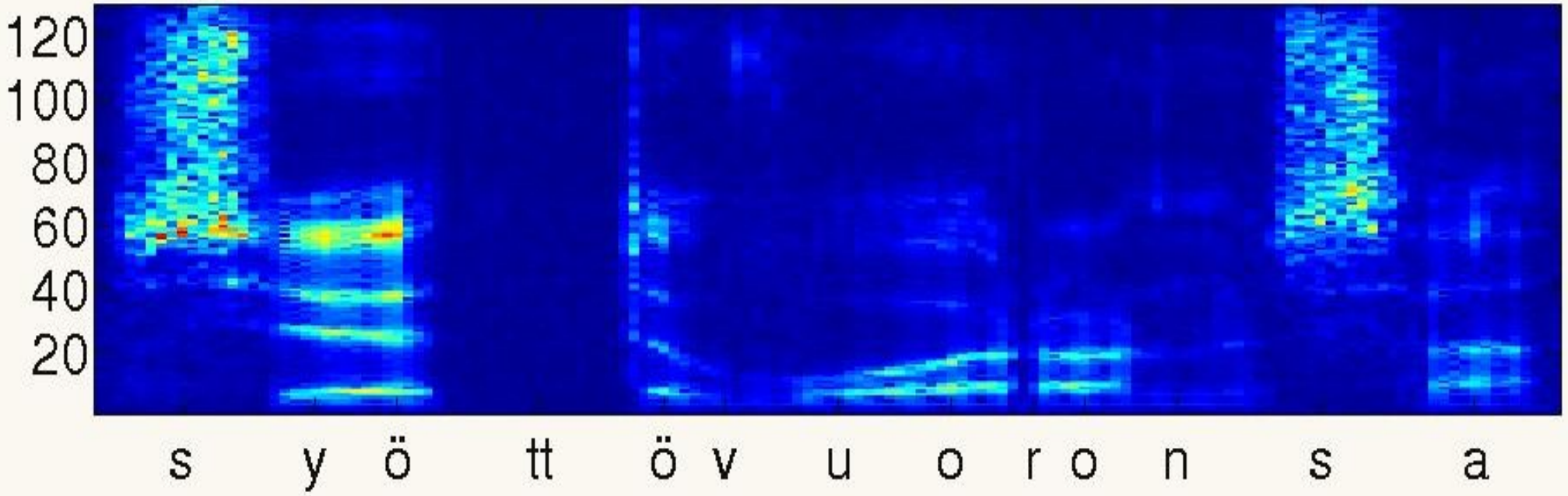
1. Learn how to collect a speech database
2. Learn what methods are used for automatic speech recognition (ASR) today
3. Learn how well the ASR systems perform in different applications

Collection of speech data

- Purposes:
  - To collect samples from over 200 speakers
  - To build a training database for speaker independent models
  - To build a large evaluation database for spoken document retrieval
  - Teaching digital processing of speech signals

Contents of the speech database

- Currently 1000 sentences read by 70 speaker
  - Digitized speech data
  - Corresponding text
  - Alignment of words in text and speech
  - Alignment of phonemes in text and speech
  - Outputs by different speech recognizers



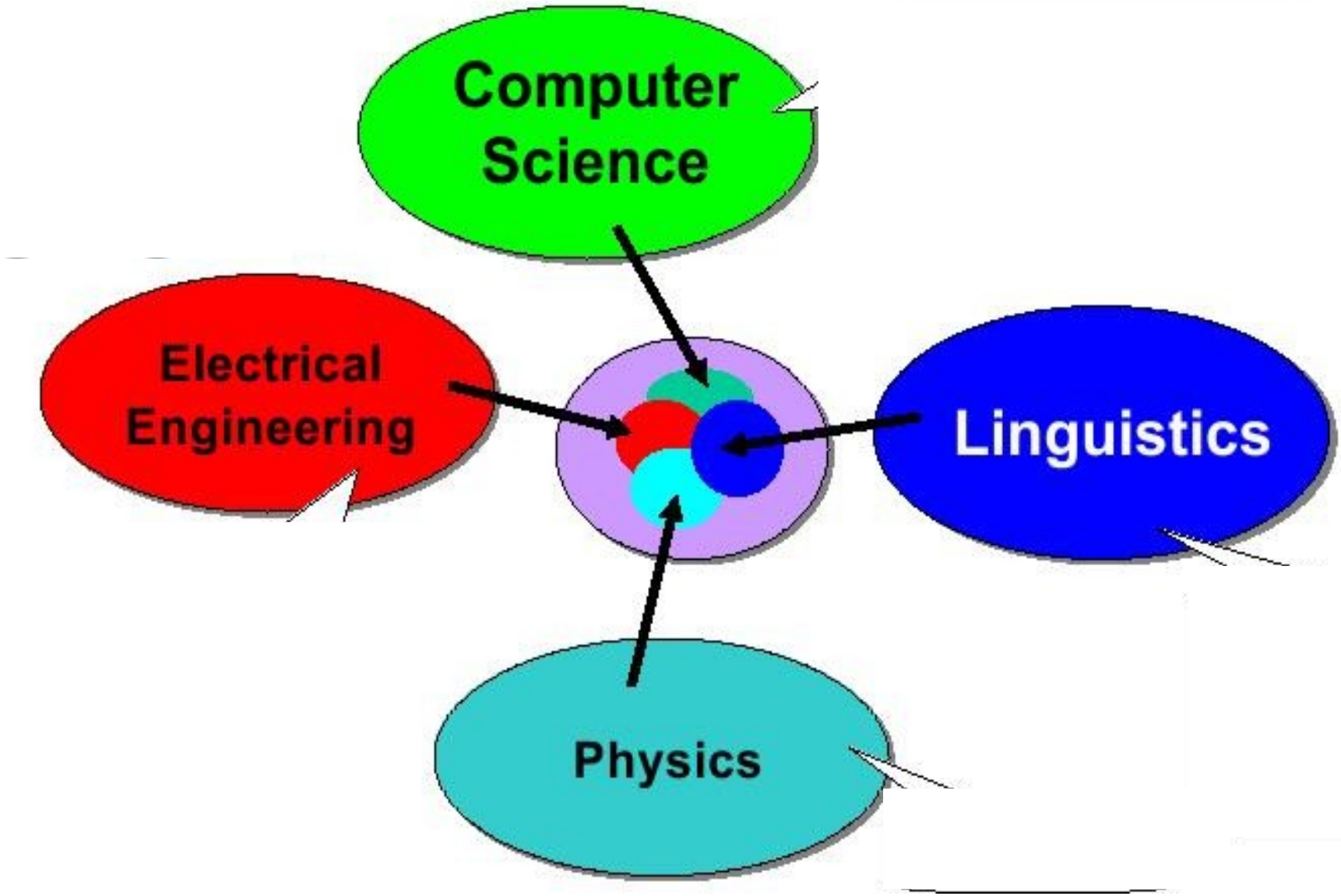
Alignment of speech and text

- Words, phonemes
- Here plotted on top of a spectrogram
  - Spectrogram = spectra in time

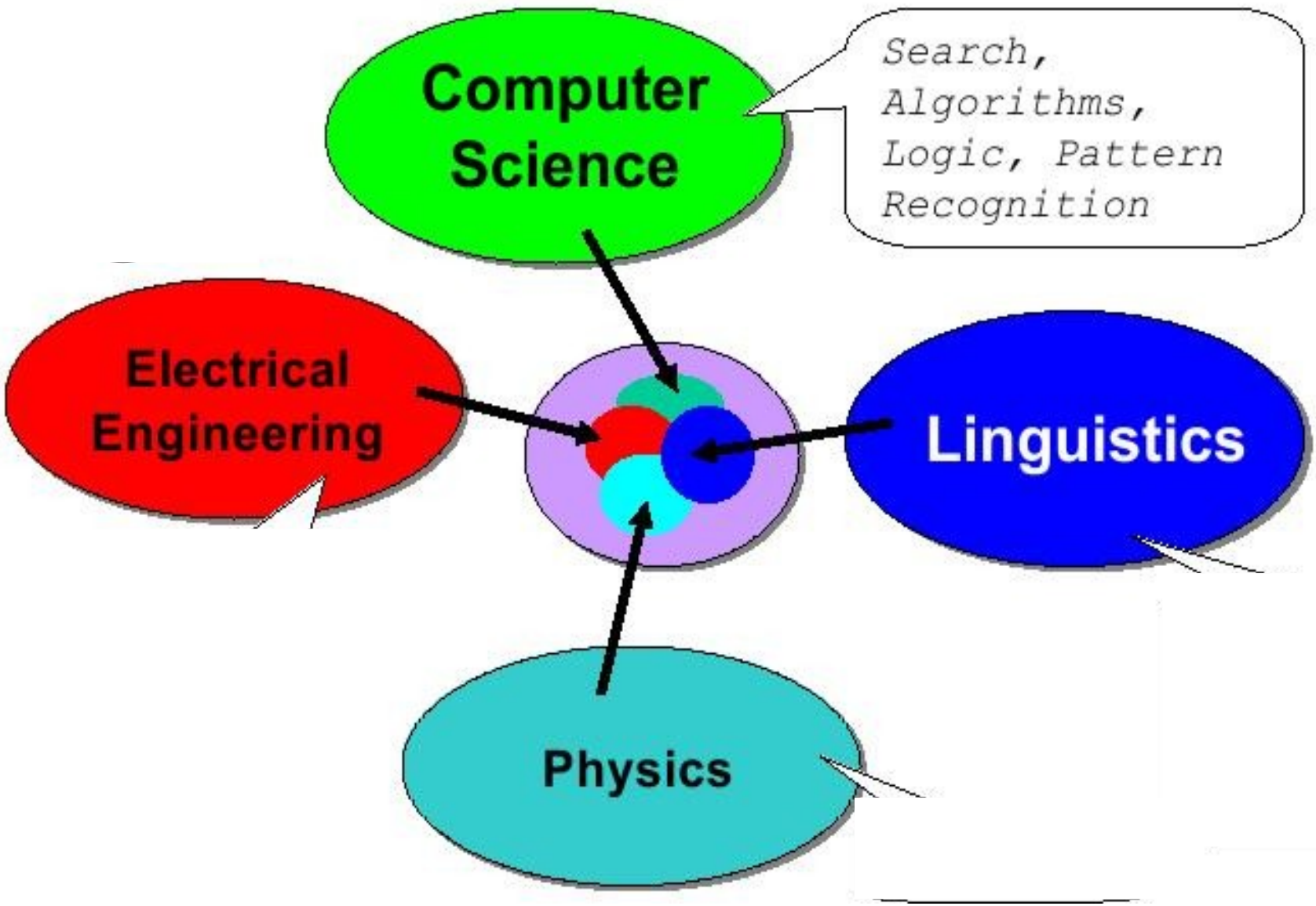
Automatic speech recognition

- Content today:
  - Collection of speech data
  - ASR systems today
  - ASR applications
  - ASR courses

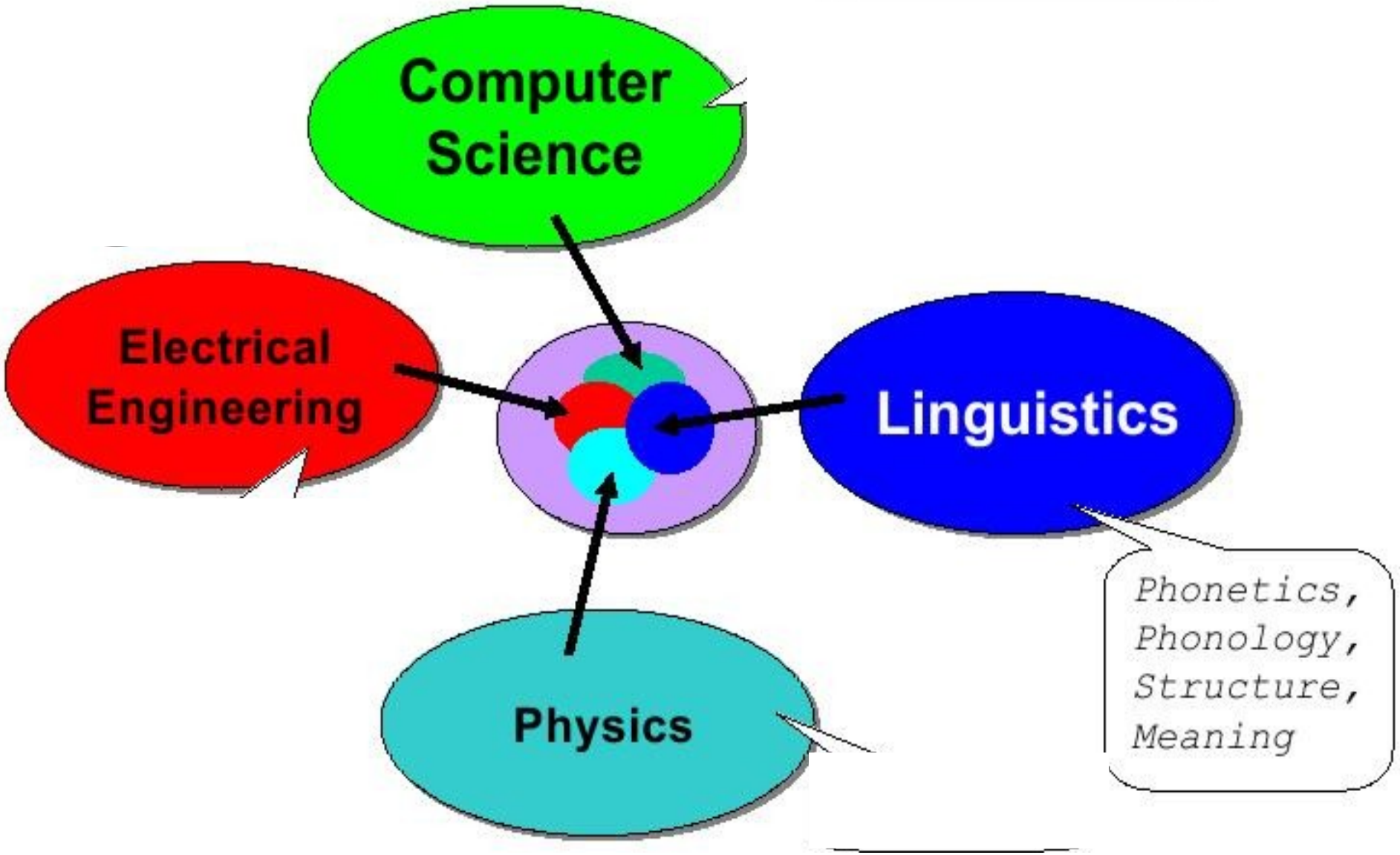
Speech recognition is multidisciplinary



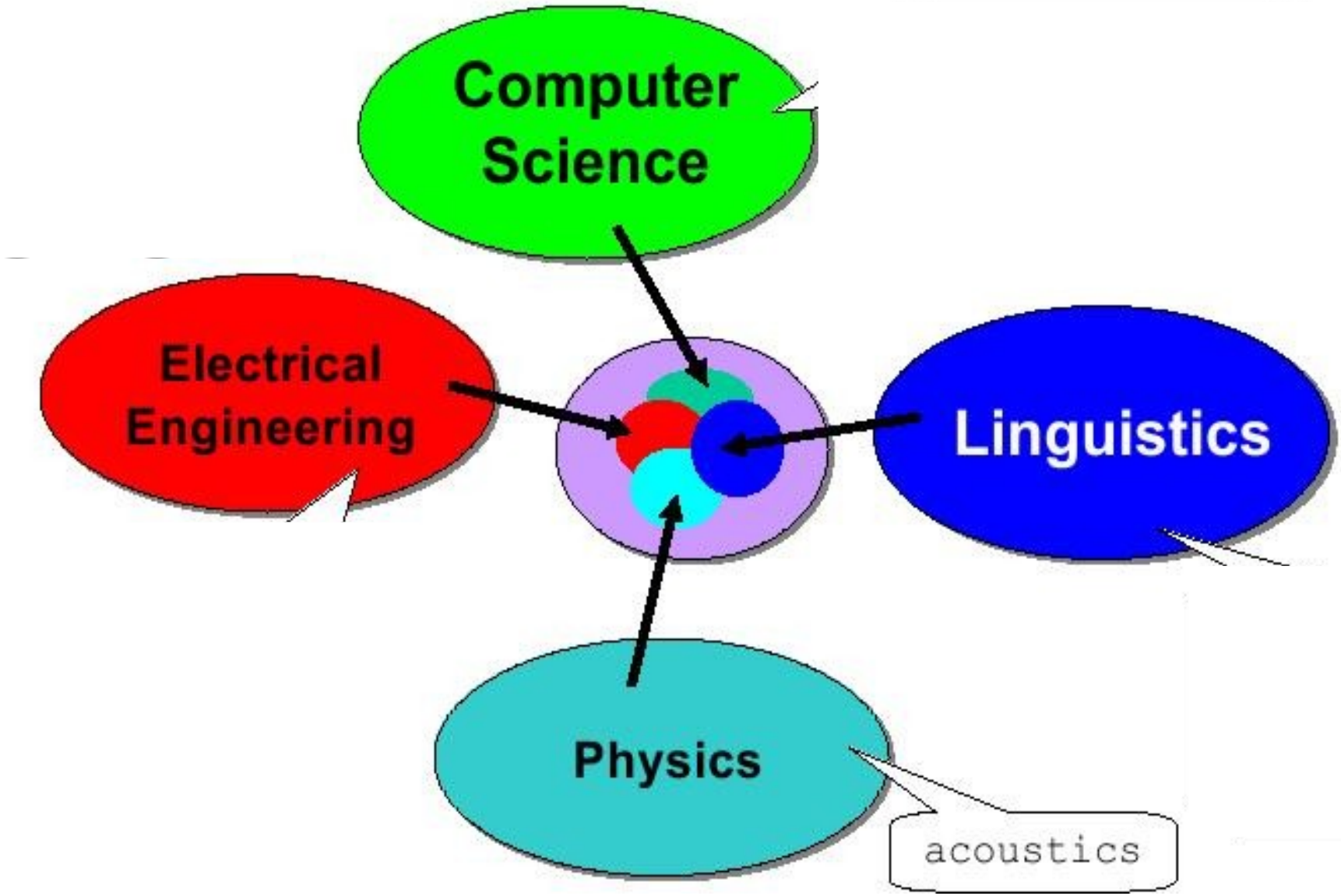
Speech recognition is multidisciplinary



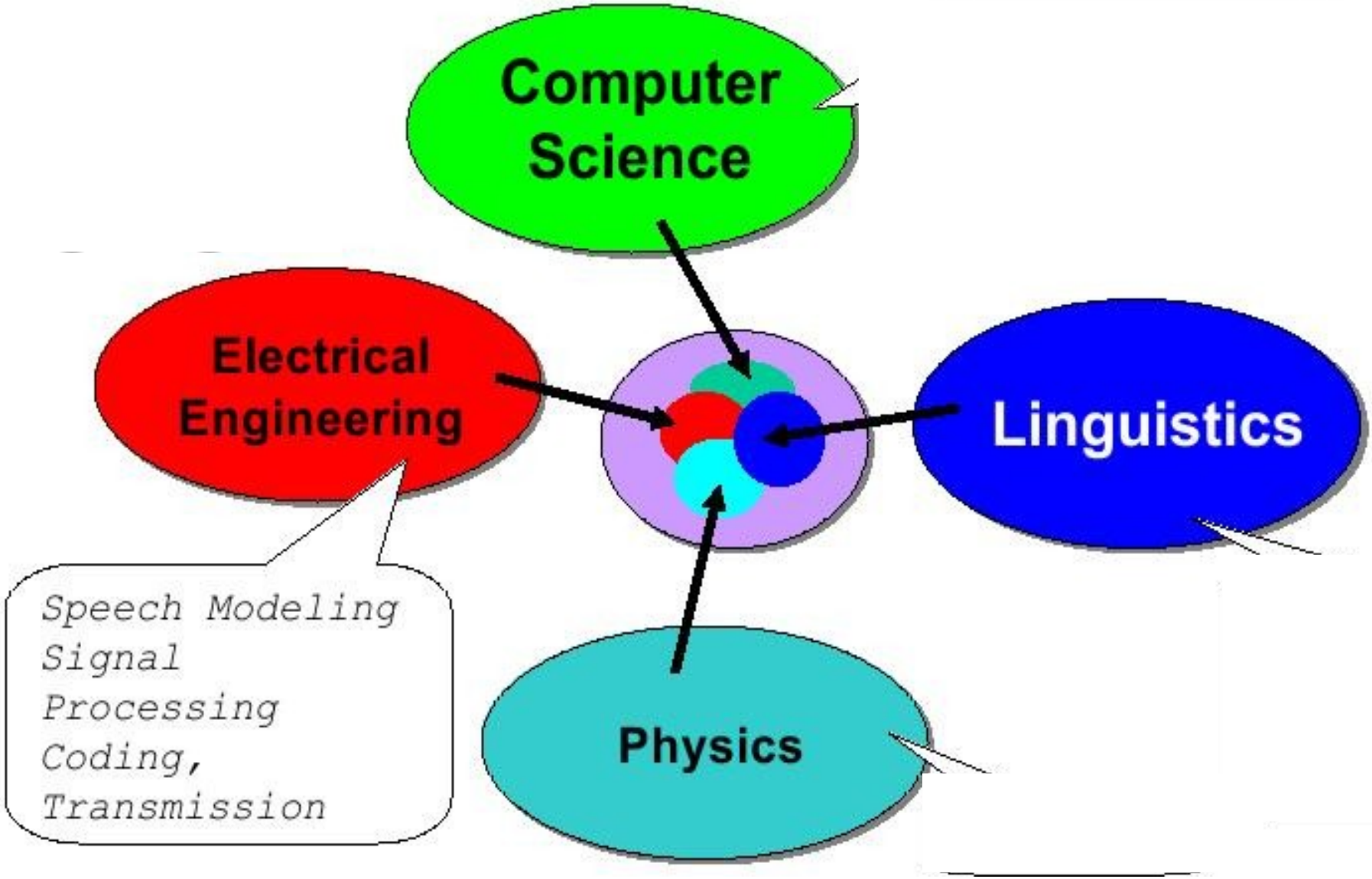
Speech recognition is multidisciplinary



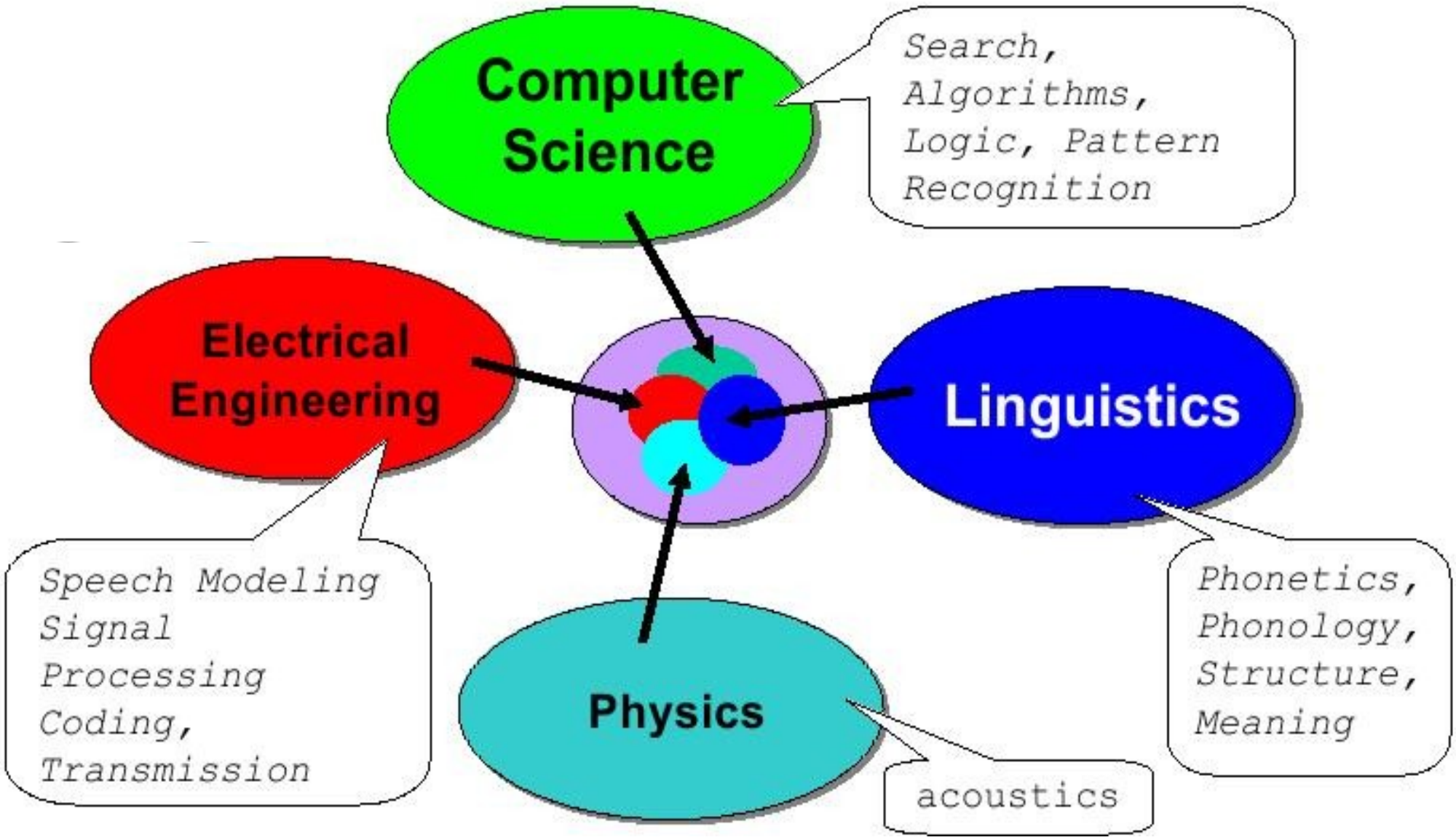
Speech recognition is multidisciplinary



Speech recognition is multidisciplinary



All these are needed to build ASR systems!



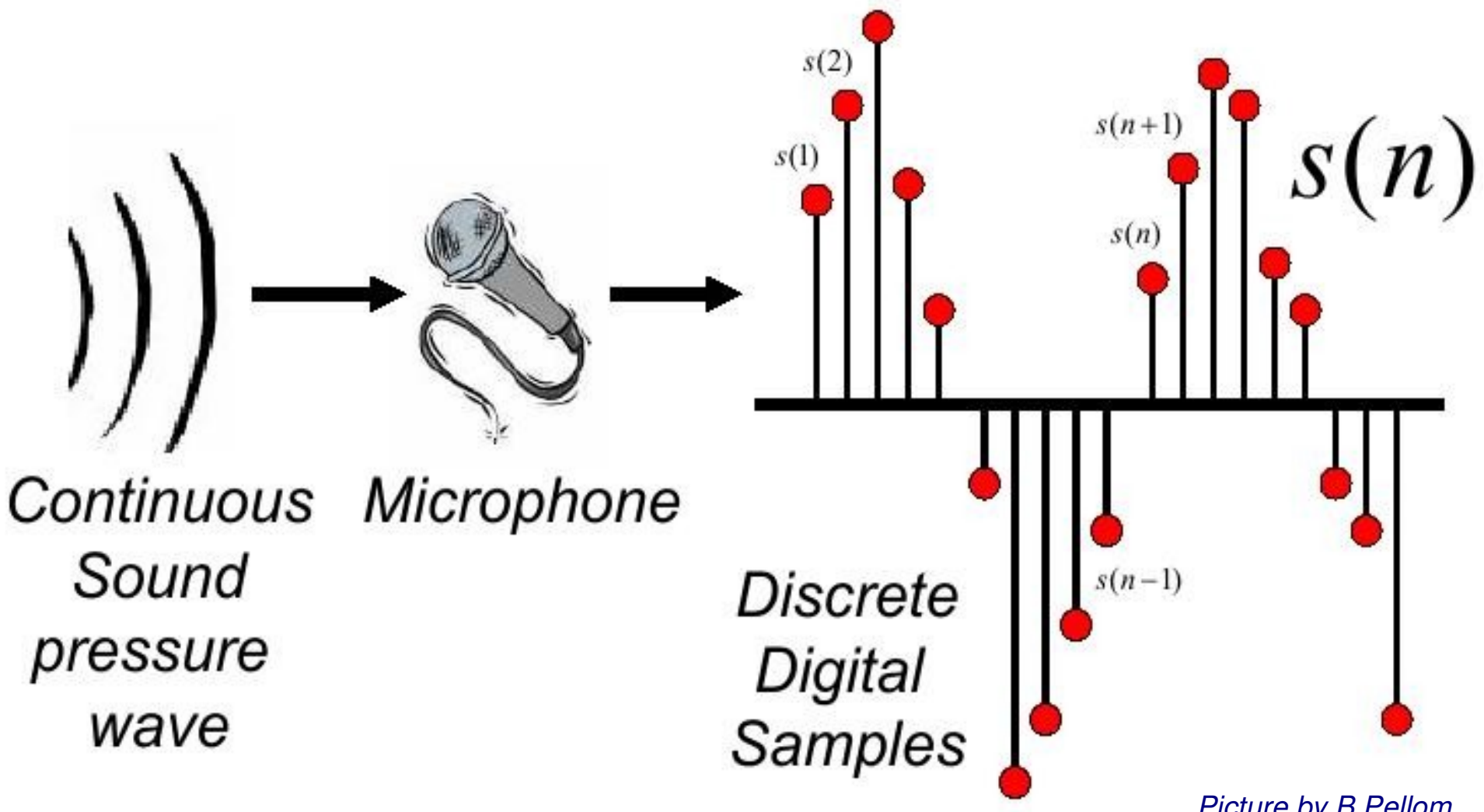
What is speech recognition?

- **Find the most likely word or word sequence given the acoustic signal and our statistical models!**
- **Language model** defines words and how likely they occur together
- **Lexicon** (vocabulary) defines the word set and how the words are formed from sound units
- **Acoustic model** defines the sound units independent of speaker and recording conditions

Automatic Speech Recognition

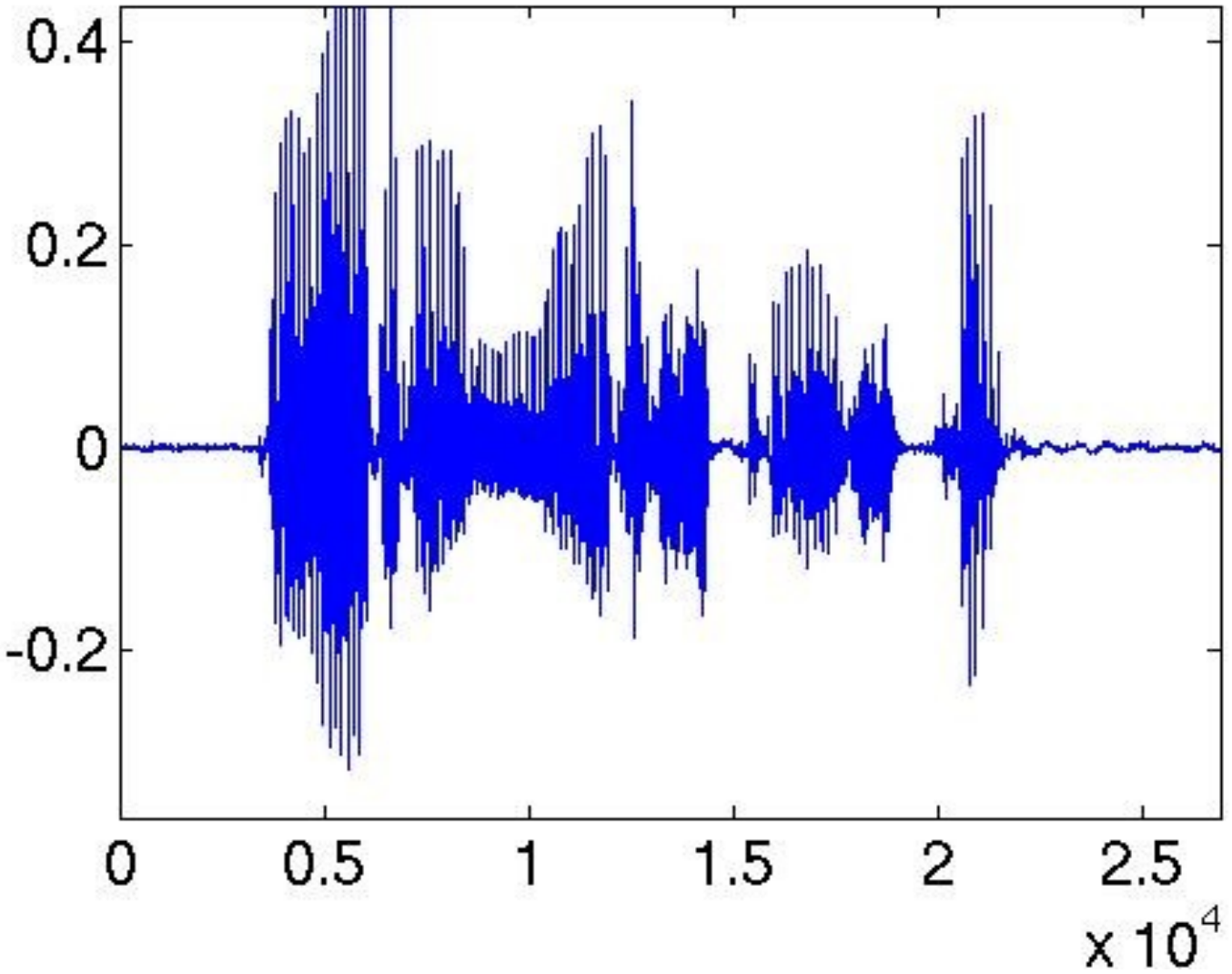
1. **Speech signal**
2. Phoneme model
3. Lexicon, Pronunciation dictionary
4. Language model
5. Recognizer structure
6. Main applications

Speech recording



Picture by B.Pellom

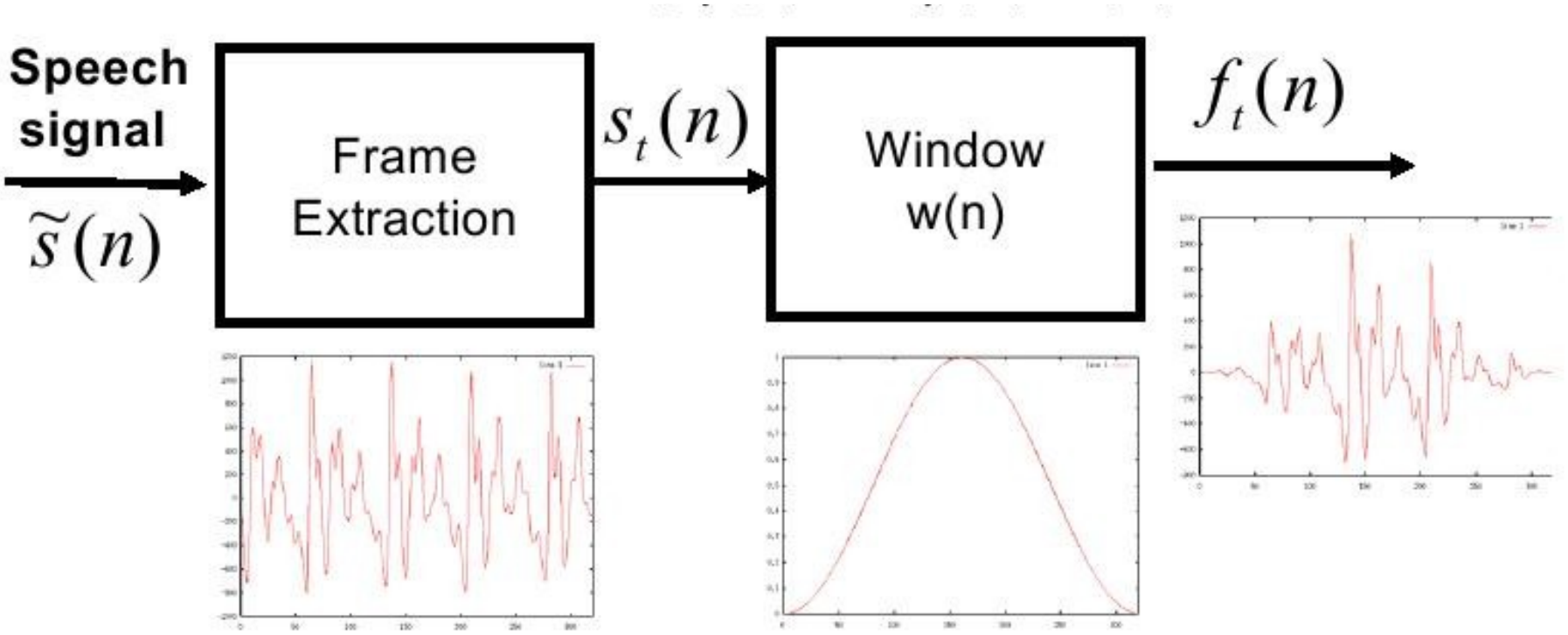
Raw speech waveform data



How to recognize speech?

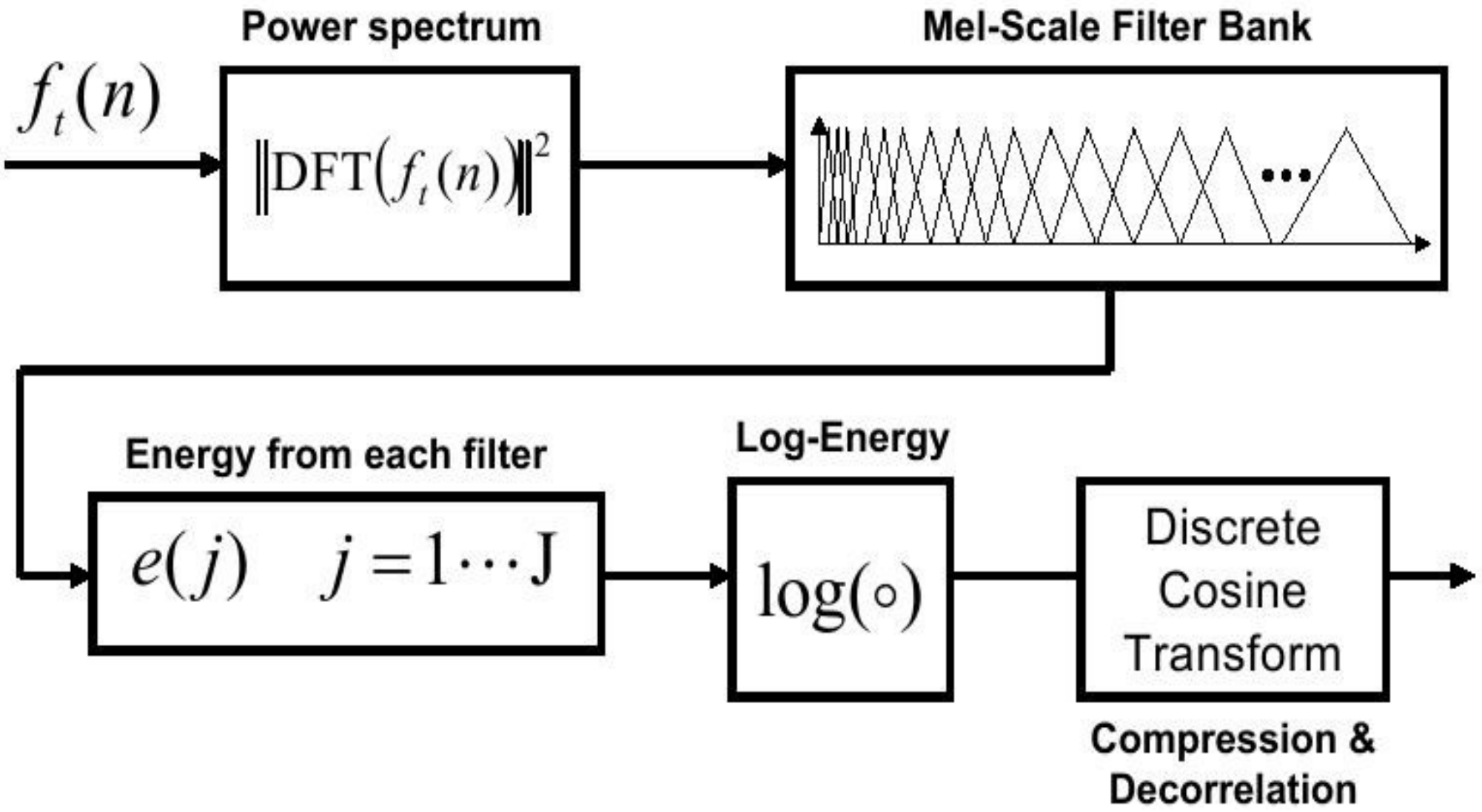
- Measure characteristic features of the signal, train models and compare them.
- Good features should be:
  - Compact
  - Discriminative for speech sounds
  - Fast to compute
  - Robust for noise

Analyse in short frames

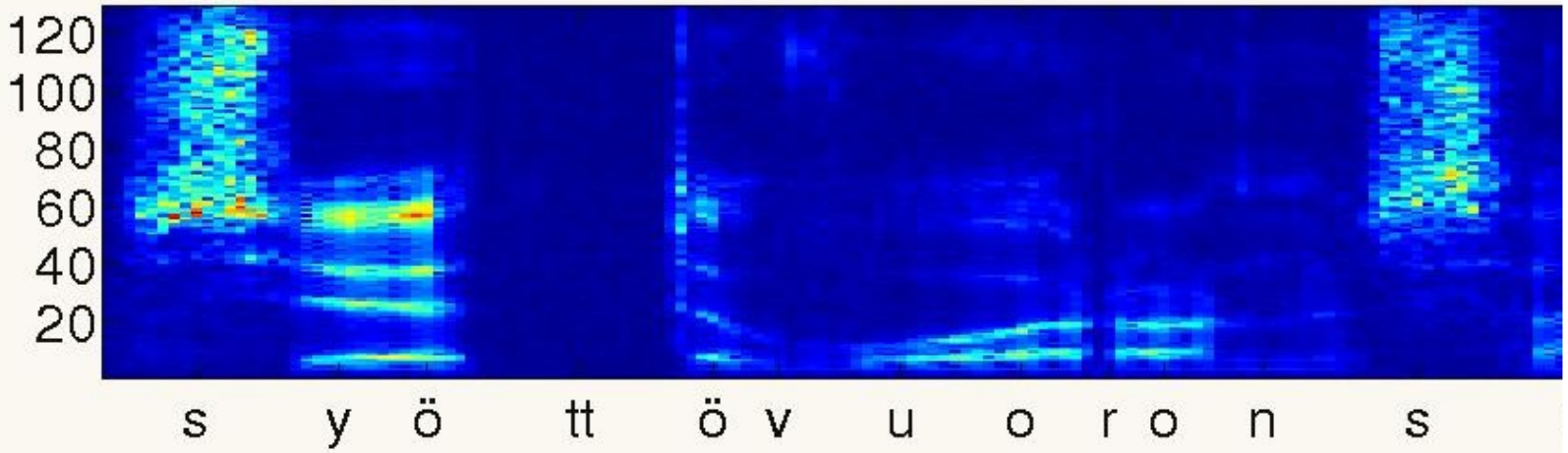


Picture by B.Pellom

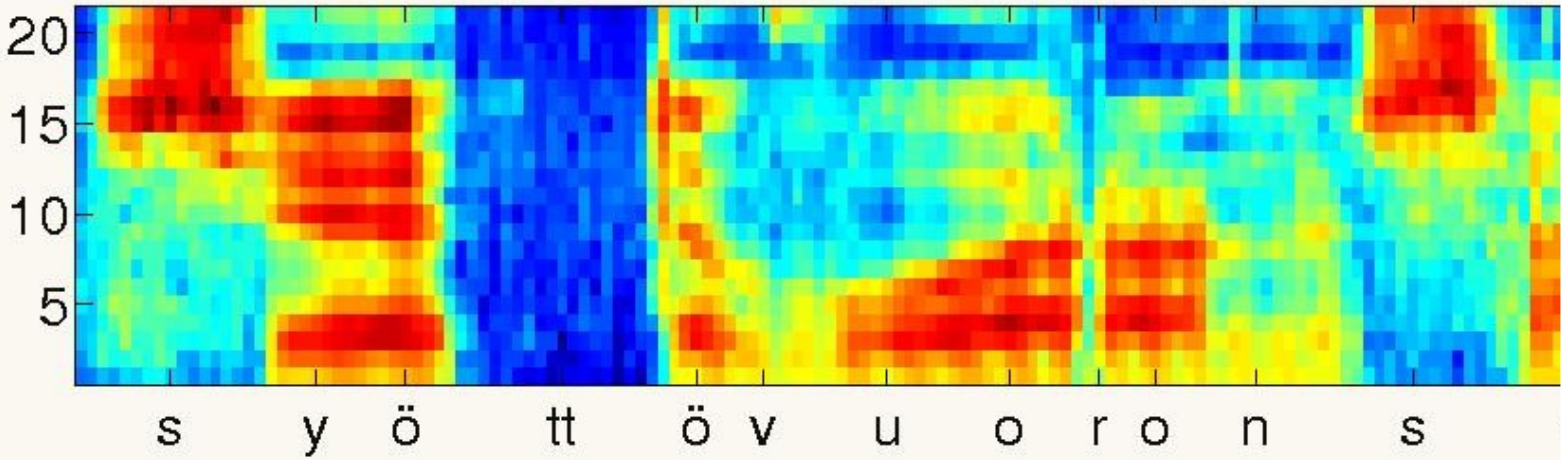
Compute features of each frame



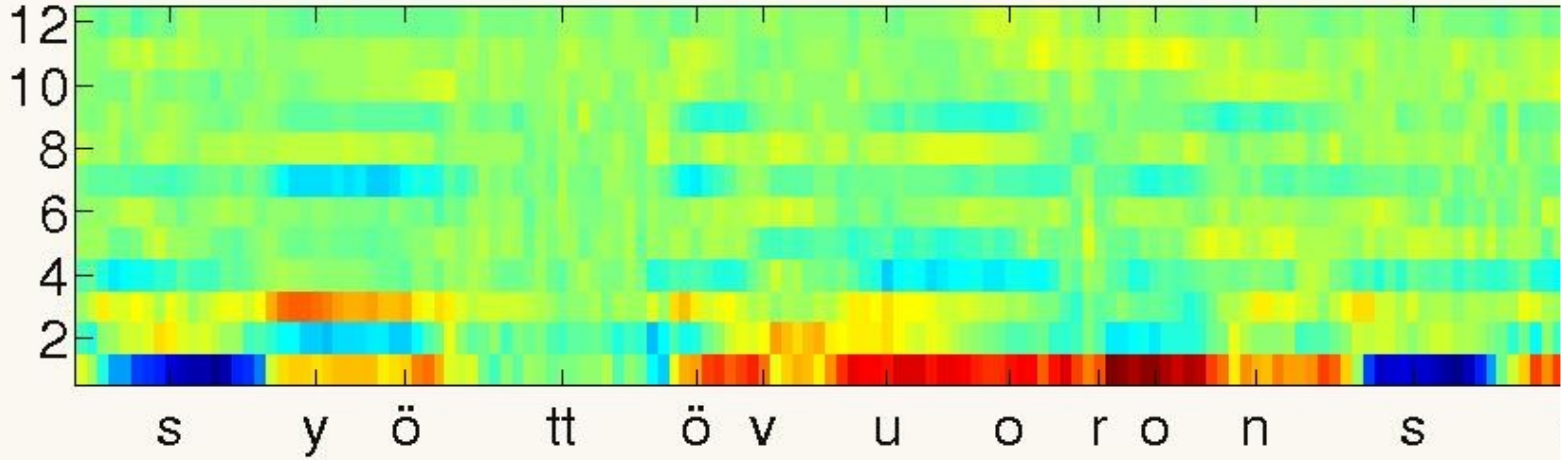
Picture by B.Pellom



1. Frames:
short 10ms
windows
2. FFT:
power spectrum
spectrogram



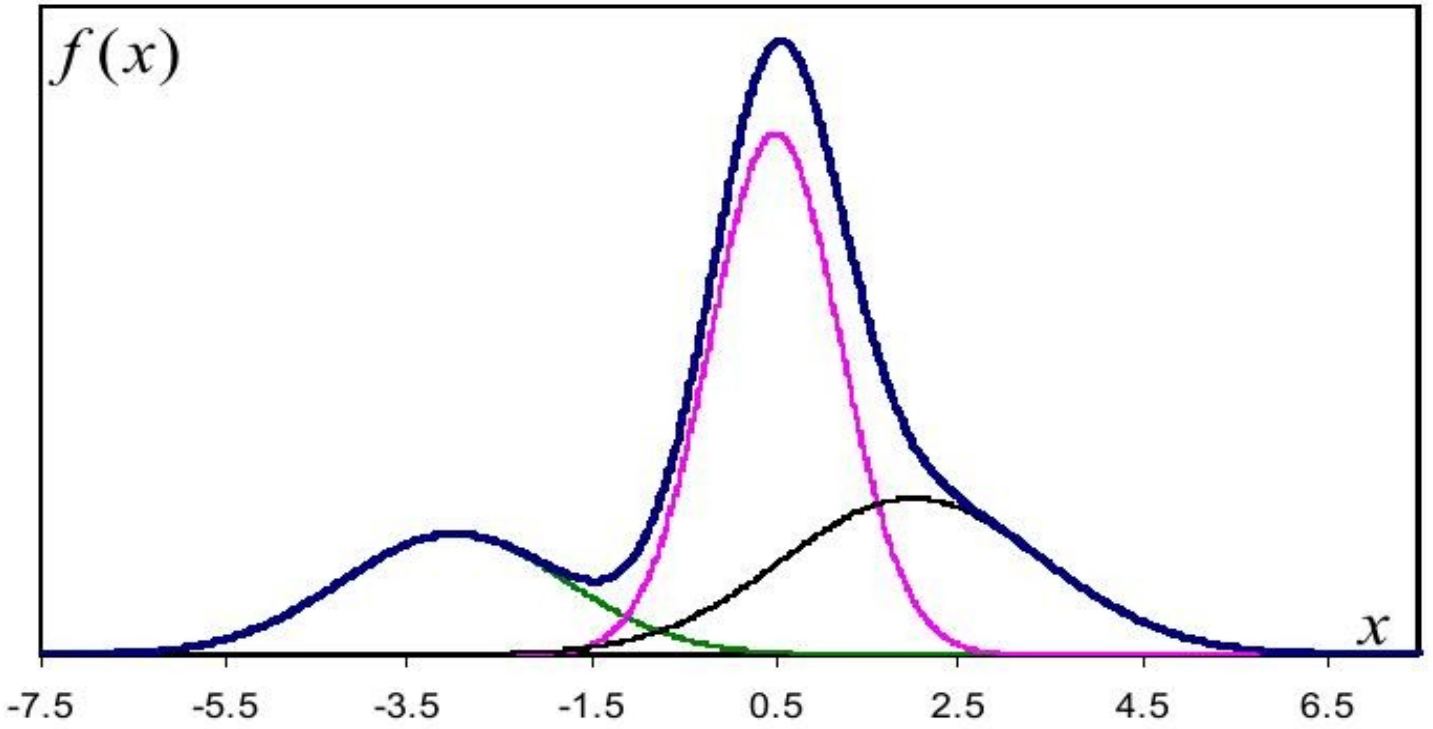
3. Filtering:
mel filter
motivated by
human ear
"essential data"



4. Features:
DCT transform
mel cepstrum
MFCC
-less features
-less correlation

Using the features to classify sounds

- Train a statistical model (mean and variance) for samples of each sound (typically a Gaussian Mixture model, GMM)
- Classify new samples by choosing the best-matching statistical model



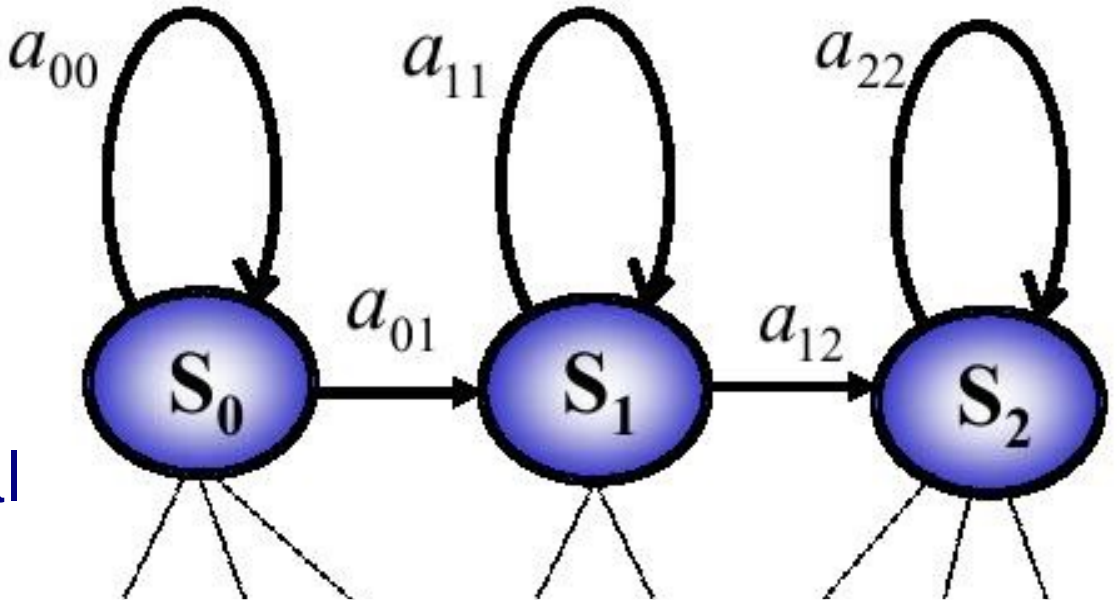
Picture by B.Pellom

Automatic Speech Recognition

1. Speech signal
2. **Phoneme model**
3. Lexicon, Pronunciation dictionary
4. Language model
5. Recognizer structure
6. Main applications

Sequence of states

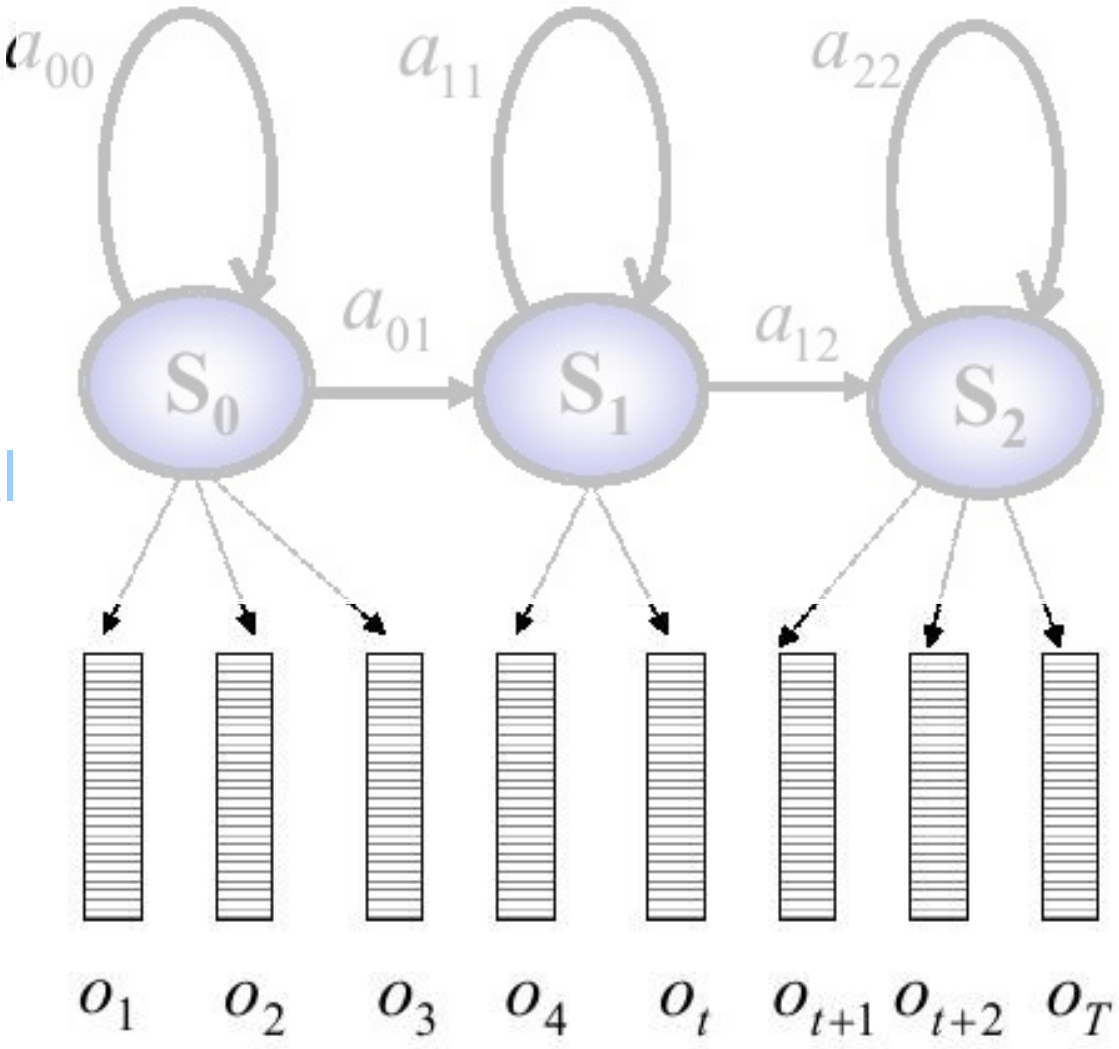
1. A system that has a set of operational states
2. From state *i* it moves to state *j* by probability *a(ij)*
3. Each state emits a characteristic sound signal
4. Signals are measured by feature vectors
5. The system's internal **state is hidden**, only the **feature vectors** can be **measured**



Picture by B.Pellom

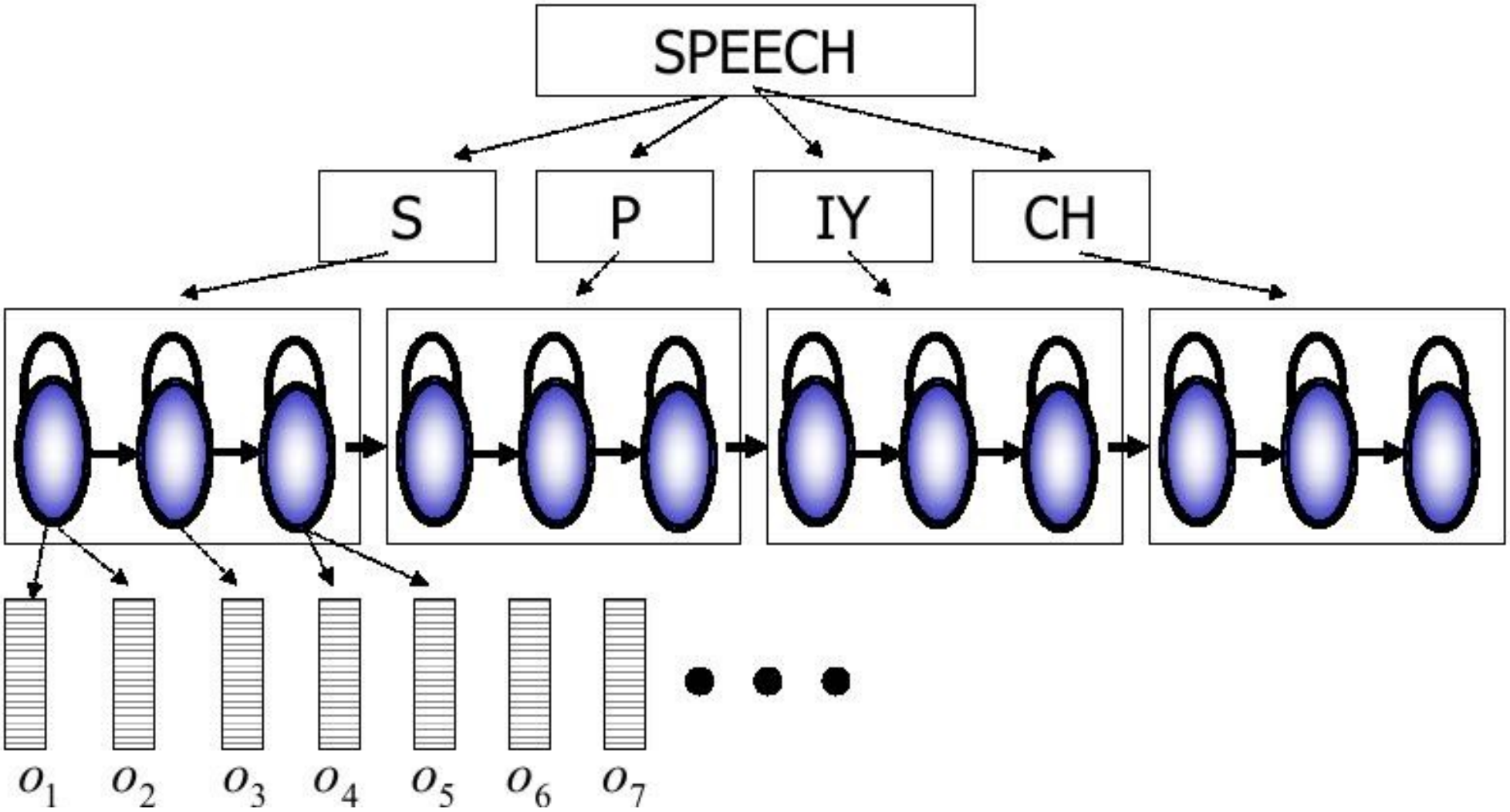
Sequence of states

1. A system that has a set of operational states
2. From state *i* it moves to state *j* by probability *a(ij)*
3. Each state emits a characteristic sound signal
4. Signals are measured by feature vectors
5. The system's internal **state is hidden**, only the **feature vectors** can be **measured**



Picture by B.Pellom

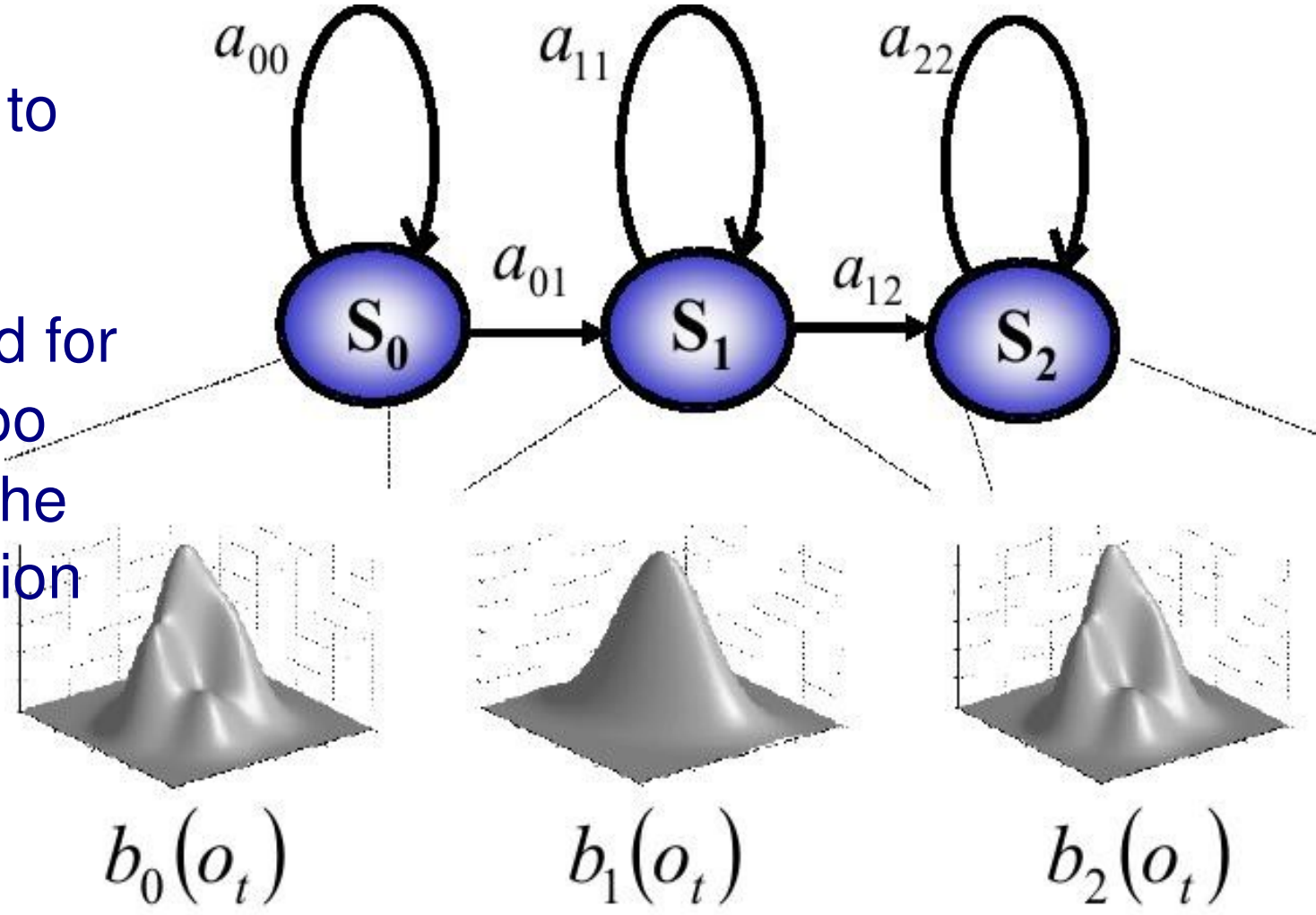
Model of words and phonemes



Picture by B.Pellom

The full model

- Each state emits sounds according to its GMM model
- This generative model can be used for **text-to-speech**, too
- The higher *a(ii)*, the longer is the duration



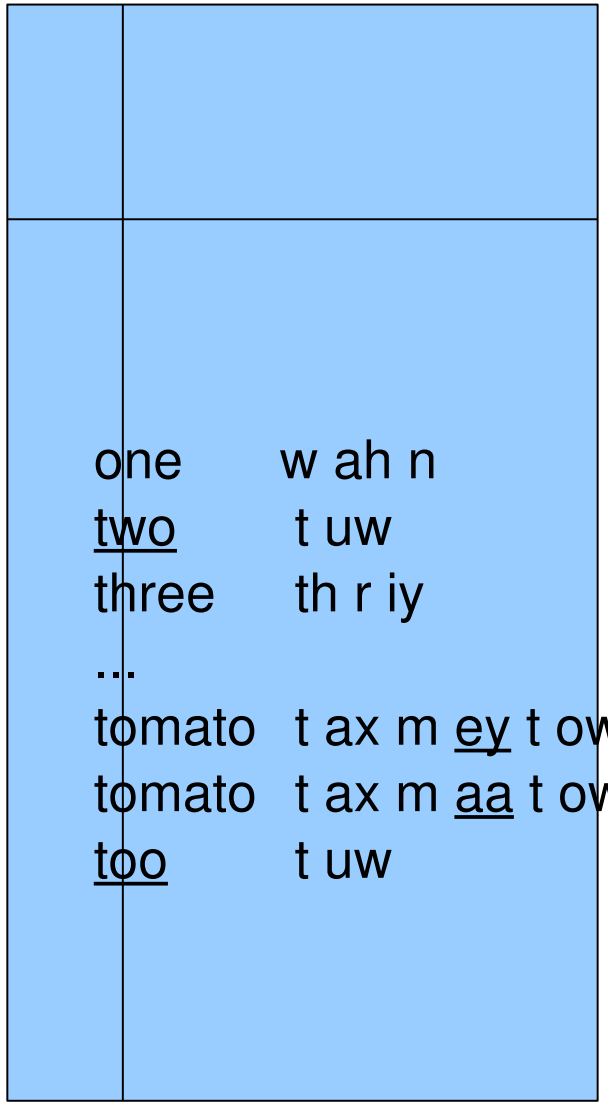
Picture by B.Pellom

Automatic Speech Recognition

1. Speech signal
2. Phoneme model
3. **Lexicon, Pronunciation dictionary**
4. **Language model, N-gram**
5. Recognizer structure
6. Main applications

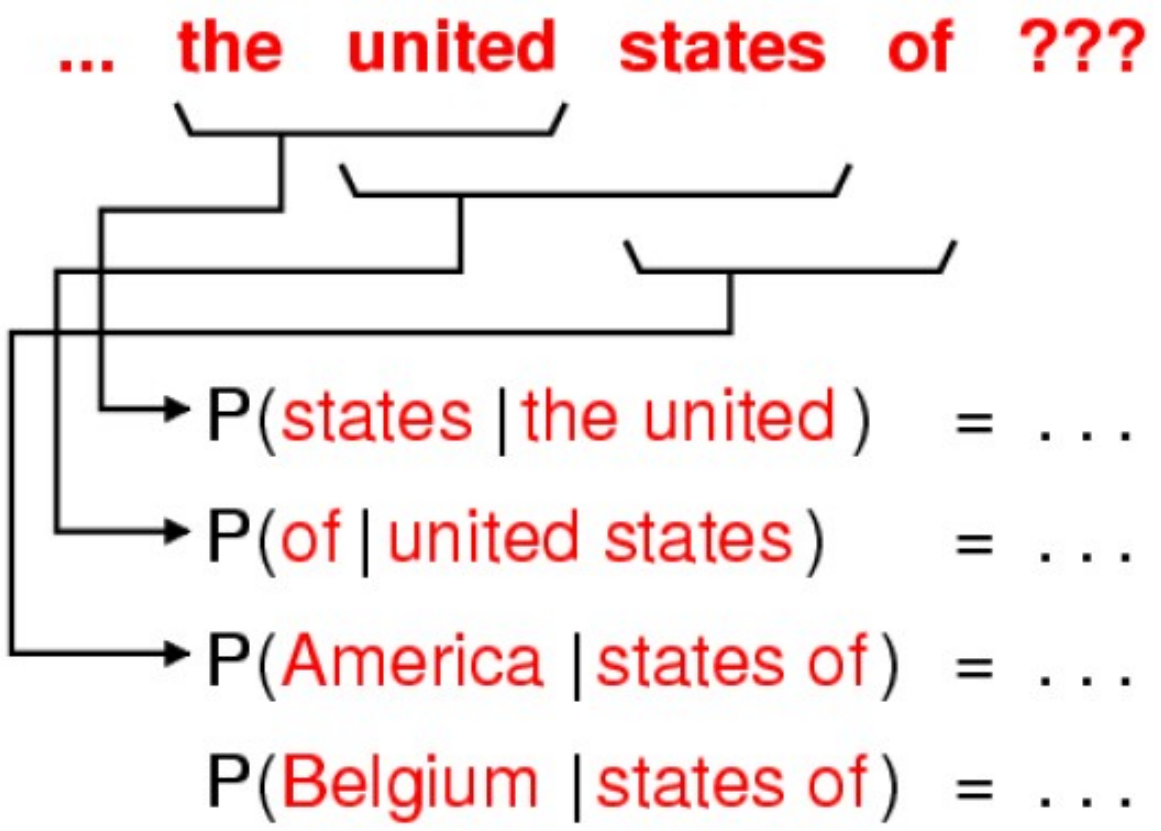
Pronunciation of words

- A lexicon or pronunciation dictionary tells how the words are pronounced
- Each word is described as a phoneme sequence
- Manually generated or using pronunciation rules, if available
- One word may have several pronunciations (with priors)
- Several words may have the same pronunciation



Language models

- trigram = 3-gram:
- Word occurrence depends only on its immediate short context
- A conditional probability of word given its context
- Estimated from a large text corpus (count the contexts!)

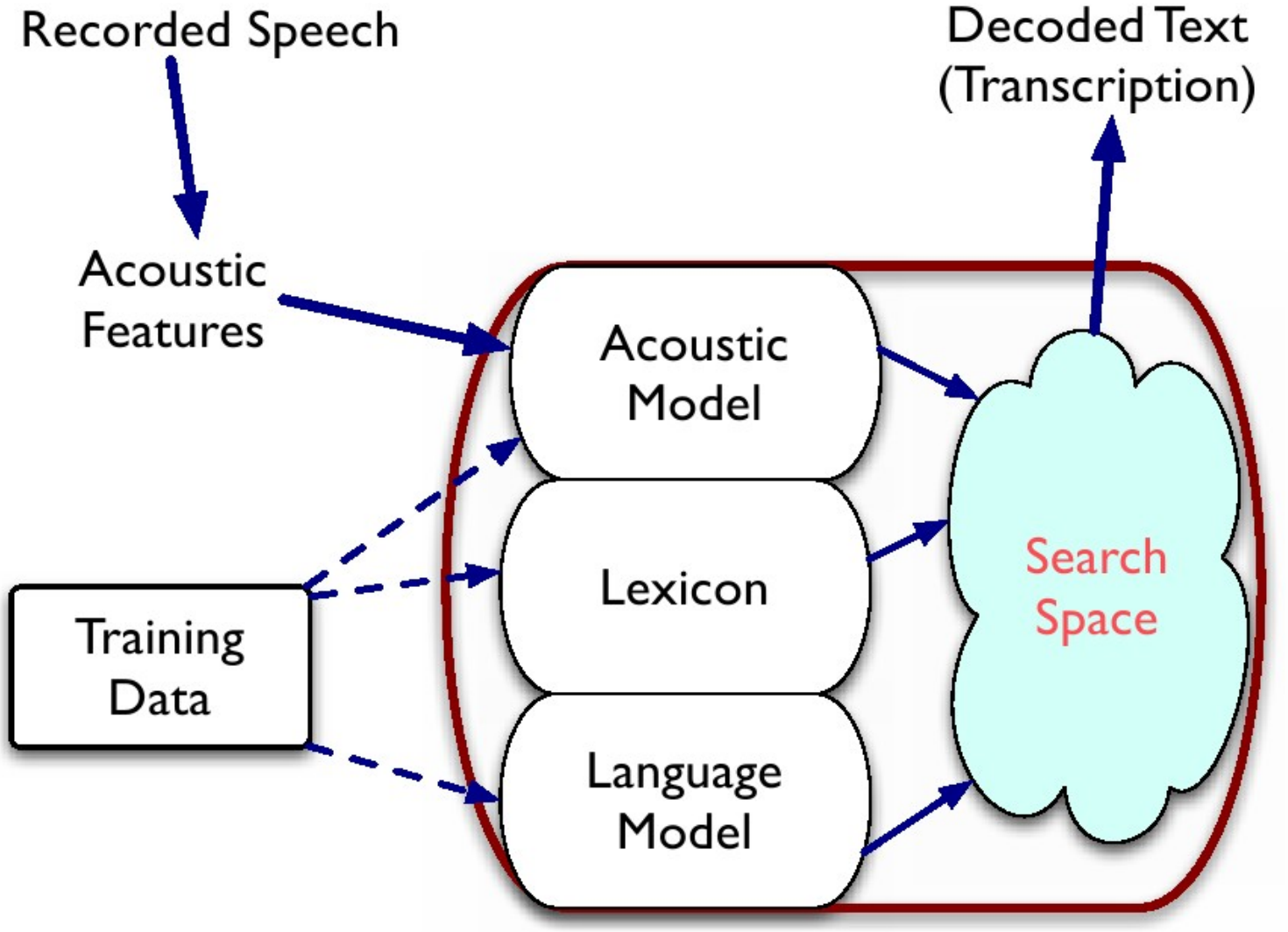


Picture by B.Pellom

Automatic Speech Recognition

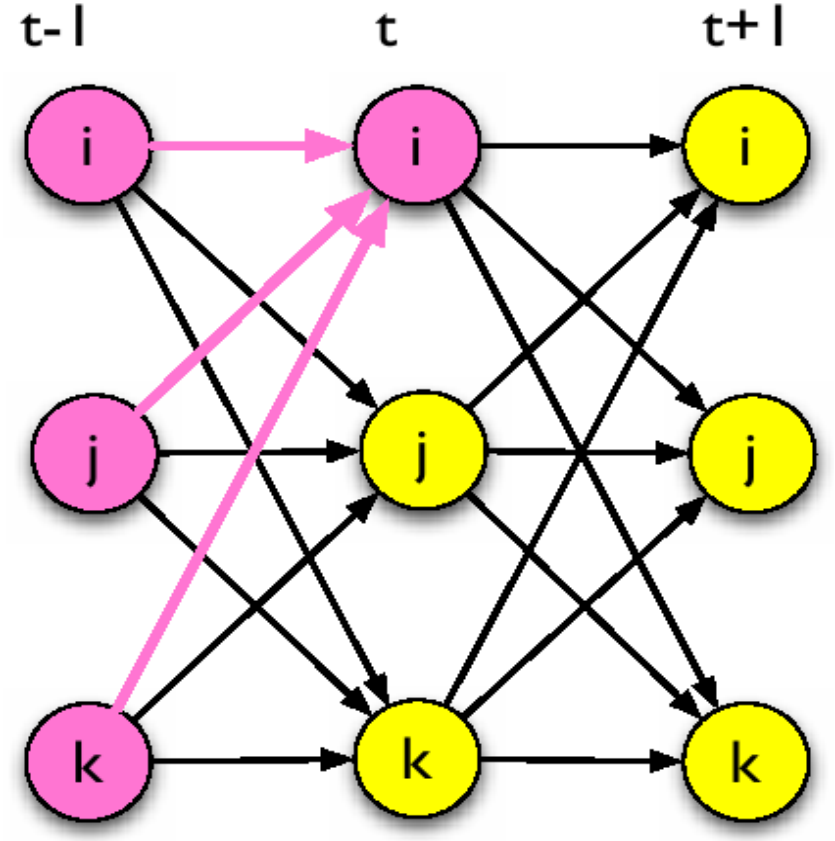
1. Speech signal
2. Phoneme model
3. Lexicon, Pronunciation dictionary
4. Language model
5. **Recognizer structure**
6. **Main applications**

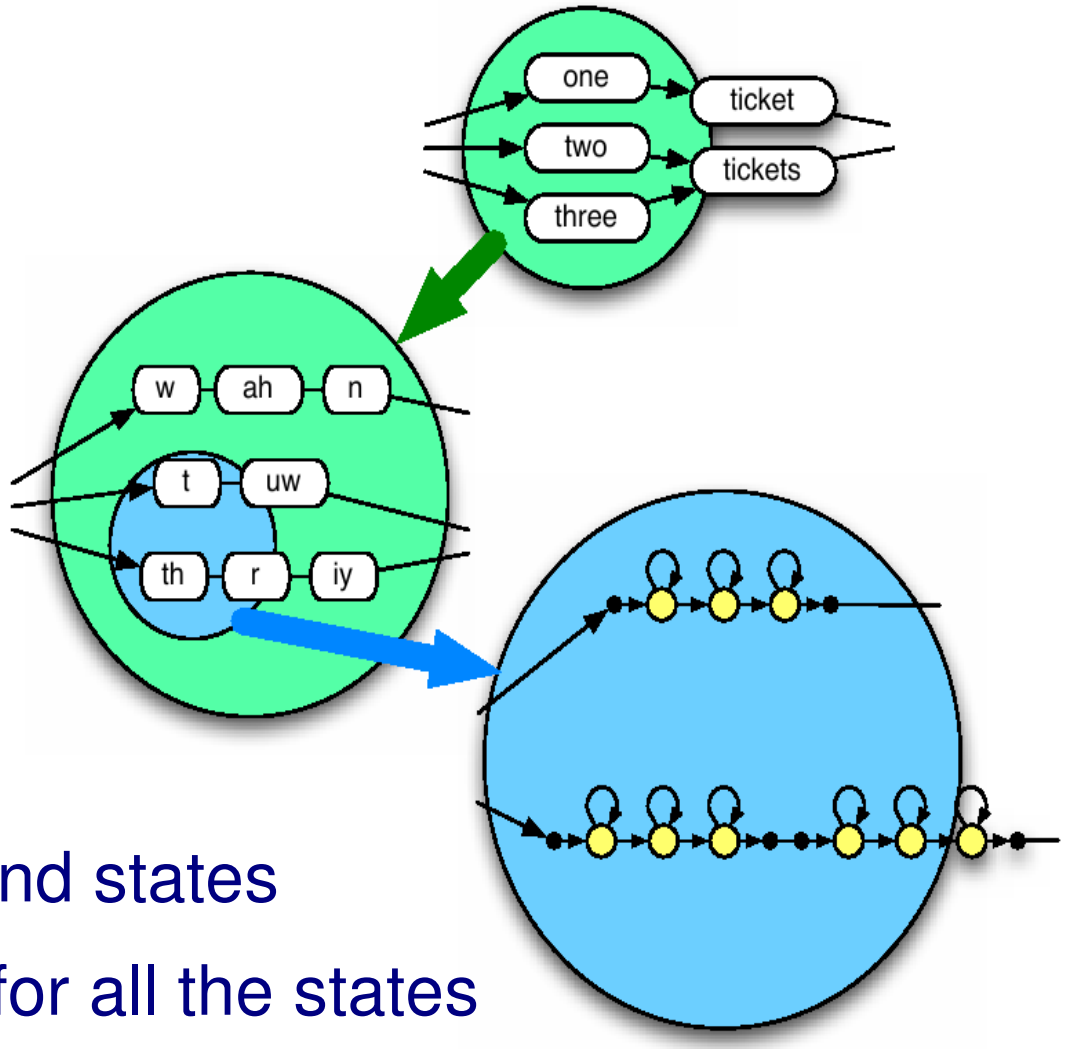
Decoding and search



Picture by S.Renals

Recognition of continuous speech





1. Map words into phonemes and states
2. Construct a search network for all the states

Picture by S.Renals

ASR tasks and solutions

- **Speaking environment and microphone**
  - Office, headset or close-talking
  - Telephone speech, mobile
  - Noise, outside, microphone far away
- **Style of speaking**

- **Speaker modeling**

ASR tasks and solutions

- **Speaking environment and microphone**
- **Style of speaking**
  - Isolated words
  - Connected words, small vocabulary
  - Word spotting in fluent speech
  - Continuous speech, large vocabulary
  - Spontaneous speech, ungrammatical
- **Speaker modeling**

ASR tasks and solutions

- **Speaking environment and microphone**
- **Style of speaking**
- **Speaker modeling**
  - Speaker-dependent models
  - Speaker-independent, average speaker models
  - Speaker adaptation

Typical applications

1. User interface by speech
2. Dictation
3. Speech translation
4. Audio information retrieval

1. User interface by speech

- Give **spoken commands** to a system
- **Feedback** often visual or synthesized speech
- Typical devices to control
  - mobile phones
  - car navigation system
  - telephone based information systems

2. Dictation

- **Online** dictation of documents, emails or SMS
  - "Speech-To-Text"
- **Offline** processing of audio files, voice mails, interviews, meeting minutes
- Goal: Get as **accurate transcription** of the input speech as possible
- Typically **the speaker knows** this and speaks **clearly and slowly** and goes into a **quiet environment**

TKK demo 1

- **Offline** dictation of fluent speech
- www.cis.hut.fi/projects/speech/demo
- **Submit an audio file and receive the transcription** by email
- The current Finnish version works best on news reading style (trained by newspapers, journals and books)
- May take a while (5 min), because these jobs run on low priority at our server

Speech recognition demo

This is a www-interface for the Finnish large vocabulary continuous speech recognizer at TKK.
Your speech file will be recognized on the laboratory's grid computing system and the results sent to you by e-mail.
If you are interested in trying the system, please ask for a password from **asrdemo@cis.hut.fi**.
For some thoughts about the system and performance, see the **FAQ**.
We reserve the right to use your file for research purposes.

Audio file:

Your audio file should contain clearly pronunciated Finnish speech.
The file should be in a format automatically recognizable by the **Sox converter**.
Waveform 16khz, 16-bit mono will give the best results.
Maximum filesize is 20 megabytes and there is an upper limit for audio files in the queue.

Browse...

Audio file sample rate:

Default is 16 kHz, but if your file is in 8kHz, check it instead.

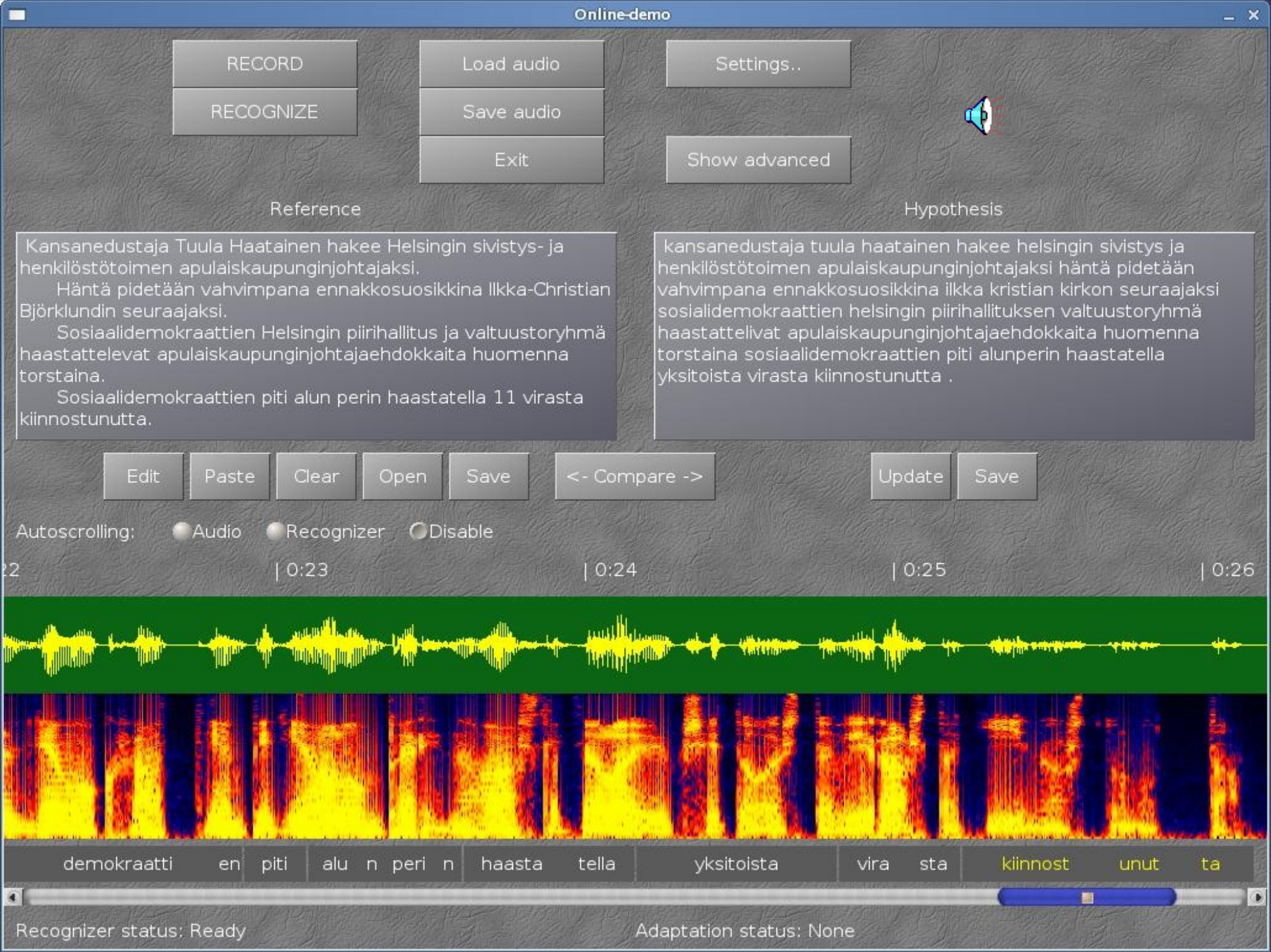
- ○ 8 kHz
- ◉ 16 kHz

E-mail address:

The E-mail address to which the results should be sent

TKK demo 2

- **Online** dictation of fluent speech
- Runs on linux workstations and laptops
- Option to run the recognition engine on a separate server or in the local machine
- The same system as in the offline www-demo, almost real-time
- Versions for Finnish, News English, Conversation English


Online-demo
RECORD
RECOGNIZE
Load audio
Save audio
Exit
Settings..
Show advanced
Reference
Kansanedustaja Tuula Haatainen hakee Helsingin sivistys- ja henkilöstötoimen apulaiskaupunginjohtajaksi.
Häntä pidetään vahvimpana ennakkosuosikkina Ilkka-Christian Björklundin seuraajaksi.
Sosiaalidemokraattien Helsingin piirihallitus ja valtuustoryhmä haastattelevat apulaiskaupunginjohtajaehdokkaita huomenna torstaina.
Sosiaalidemokraattien piti alun perin haastatella 11 virasta kiinnostunutta.
Hypothesis
kansanedustaja tuula haatainen hakee helsingin sivistys ja henkilöstötoimen apulaiskaupunginjohtajaksi häntä pidetään vahvimpana ennakkosuosikkina ilkka kristian kirkon seuraajaksi sosialidemokraattien helsingin piirihallituksen valtuustoryhmä haastattelivat apulaiskaupunginjohtajaehdokkaita huomenna torstaina sosiaalidemokraattien piti alunperin haastatella yksitoista virasta kiinnostunutta .
Edit
Paste
Clear
Open
Save
<- Compare ->
Update
Save
Autoscrolling:
Audio
Recognizer
Disable
| 0:23
| 0:24
| 0:25
| 0:26
demokraatti en piti alu n peri n haasta tella yksitoista vira sta kiinnost unut ta
Recognizer status: Ready
Adaptation status: None

3. Speech translation

- Online translation of input speech to another language
- **Combination of ASR and MT** (Machine translation)
- May also include TTS (Text-To-Speech)
  - "Speech-To-Speech" translation
- The task is often limited to a **specific domain** (e.g. travel) or even only to specific phrases

TKK demo 3

- **Offline** Speech-To-Speech translation of fluent speech
- Similar web-based service as the TKK demo 1
- Submit an audio file and receive the transcript, translation, and synthesized speech by email
- The current **FIN-ENG version** works best on bible-style talk (parallel text MT training material was the Bible)
- **Simple concatenation of ASR-MT-TTS**, not very high performance

http://www.cis.hut.fi/projects/speech/translate/

LABORATORY OF COMPUTER AND INFORMATION SCIENCE
ADAPTIVE INFORMATICS RESEARCH CENTRE

CIS

Home | Contact | Research | Teaching | People | Jobs | Demos | Software

Speech to speech translation demo

This is a www-interface for the Finnish to English speech translation system at TKK.
Your audio submission will be recognized, translated and synthesized on the laboratory's grid computing system
If you are interested in trying the system, please ask for a password from asrdemo@cis.hut.fi.

Audio file:

Your audio file should contain clearly pronounced Finnish speech.
The file should be in a format automatically recognizable by the Sox converter.
Waveform 16khz, 16-bit mono should be preferred.
Maximum filesize is 20 megabytes and there is an upper limit for audio files in the queue.

Browse...

Reference text (optional):

Submit the correct text if you wish to get error rates along with the result

Reference translation (optional):

Submit the correct translation if you wish to get BLEU score along with the result

Typical applications

1. **User interface by speech**
2. **Dictation**
3. **Speech translation**
4. **Audio information retrieval**

4. Audio information retrieval

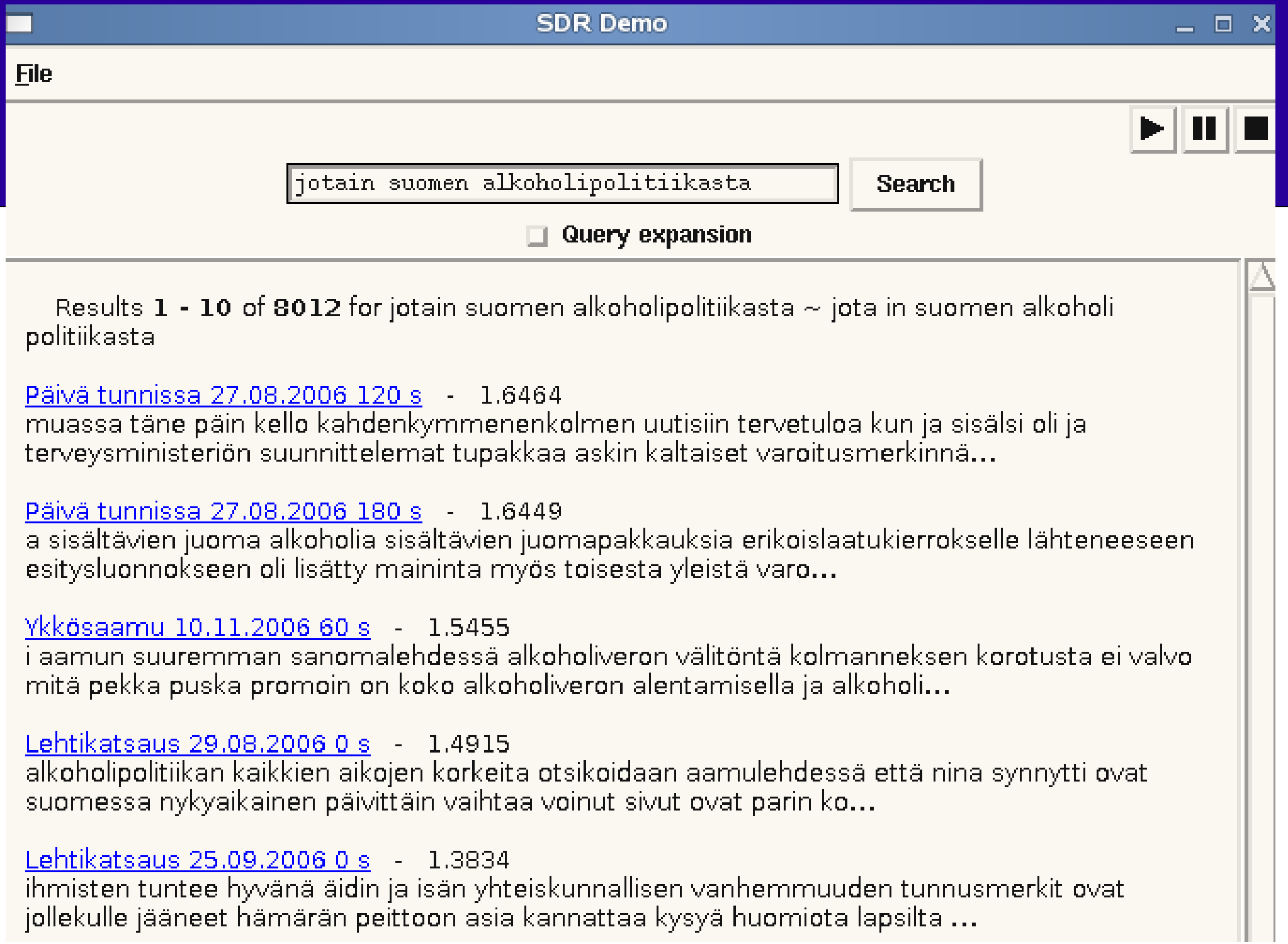
- Main goal: ASR should provide **raw text output** that contain enough correct words, **100% not required**
- Typical tasks: Spoken Document Retrieval (SDR), Audio Indexing, Speech summarization
- Typically spoken for another purpose (to human listeners), may contain **difficult and fast speech** and poor recording conditions
- Speakers may vary quickly, adaptation difficult
- Important to recognize rare words, “good indexing terms”

Indexing and retrieval

- 1. All speech **transformed to text**
- 2. The raw text output **indexed as normal text** documents
- 3. **Relevant documents ranked** for each query (as in normal search engines)
- 4. The user may **read or play** the results
- The raw text outputs much easier to browse than audio format files

TKK demo 4

- Finnish broadcast radio news recognized by TKK's LVCSR system
- The raw text output indexed by LEMUR system
  - TFIDF (Term Freq. / Document Freq.) weights
- You can compose queries and browse and listen the results using a simple IR interface


SDR Demo
File
jotain suomen alkoholipolitiikasta
Search
Query expansion
Results 1 - 10 of 8012 for jotain suomen alkoholipolitiikasta ~ jota in suomen alkoholi politiikasta
Päivä tunnissa 27.08.2006 120 s - 1.6464
muassa täne päin kello kahdenkymmenenkolmen uutisiin tervetuloa kun ja sisälsi oli ja terveysministeriön suunnittelemat tupakkaa askin kaltaiset varoitusmerkinnä...
Päivä tunnissa 27.08.2006 180 s - 1.6449
a sisältävien juoma alkoholia sisältävien juomapakkauksia erikoislaatukierrokselle lähteneeseen esitysluonnokseen oli lisätty maininta myös toisesta yleistä varo...
Ykkösaamu 10.11.2006 60 s - 1.5455
i aamun suuremman sanomalehdessä alkoholiveron välitöntä kolmanneksen korotusta ei valvo mitä pekka puska promoin on koko alkoholiveron alentamisella ja alkoholi...
Lehtikatsaus 29.08.2006 0 s - 1.4915
alkoholipolitiikan kaikkien aikojen korkeita otsikoidaan aamulehdessä että nina synnytti ovat suomessa nykyaikainen päivittäin vaihtaa voinut sivut ovat parin ko...
Lehtikatsaus 25.09.2006 0 s - 1.3834
ihmisten tuntee hyvänä äidin ja isän yhteiskunnallisen vanhemmuuden tunnusmerkit ovat jollekulle jääneet hämärän peittoon asia kannattaa kysyä huomiota lapsilta ...

jotain suomen alkoholipolitiikasta

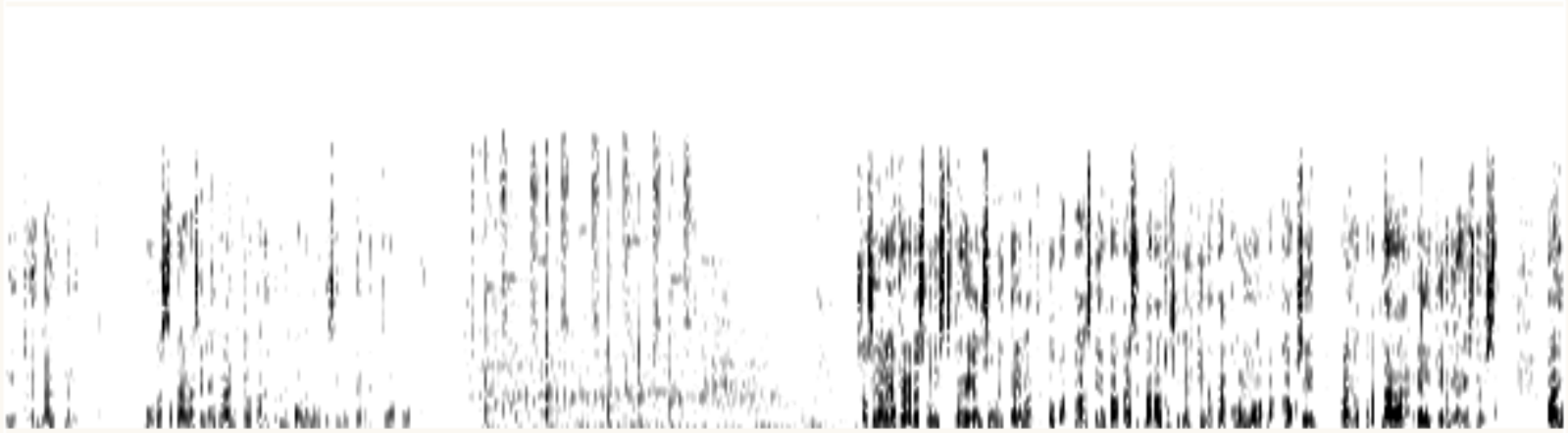
Search

Query expansion

Back to results

Päivä tunnissa 27.08.2006 120 s - Listen

muassa täne päin kello kahdenkymmenenkolmen uutisiin tervetuloa kun ja sisälsi oli ja terveysministeriön suunnittelemat tupakkaa askin kaltaiset varoitusmerkinnät alkoholijuomat pakkauksiin aiheuttavat närää elinkeinojärjestöissä että panimoliiton mukaan varoitusmerkit ovat teholtaan hyödyttää mielikuvia kustannuksiltaan kohtuuttomia peruspalveluministeri liisa hyssälä on panimoliiton kanssa täysin eri mieltä ja piinaa odotus että tuhatta kertaan uudelle pakkaus joka maksaa koko perheen ja ylimmät ottaessaan suomen ja ruotsin ole vaatinut tiedot ikä tehdä ja lehkosuo kehrä kokoon kuitenkin että kustannuskulttuurin sanoi ministeri myös on hallitus hyväksyi maaliskuussa esityksen alkoholipolitiikan muutoksista joka sisälsi muun muassa sikiövauriokartoituksen alkoholia sisältävien juoma alkoholia sisältävien juomapakkauksia erikoislaatukierrokselle lähteneeseen esitysluonnokseen oli lisätty maininta myös toisesta yleistä varoituksista alkoholin terveyshaitoista on on toimitusjohtajana timo jaatisen mukaan tämä alkaa muistuttaa liikaa amerikkalaisia käytäntöjä tämän amerikkalainen ajattelu jossa kuluttaja odotetaan mitä ihmeellisimmistä asioista tulee seminaarin pidetään kesän todetaan alkoholin esimerkiksi minkälaisia käsitys että suomalaiset kuluttajat aikaa jolloin pitäisi mitä haittavaikutuksia alkoholin tai väärinkäytöstä on s pitäisi keskittyä sellaiseen toimenpiteisiin jotka perehtyisi väärinkäyttöä ja toisaalta alaikäisten alkoholinkäyttönsä sieltä vastuullisia kuluttaja tavallista kadunmiestä syyllistetään heikoi näyteä merkkinä tosin tehokkaita koneita kaupasta saatava on ollut toimeen veloittaa markasta kahdeksan luonnosta vesistölta syyriaa tulevan vuoden on saatu noin miljoona



Mik

Automatic speech recognition

- Content today:
  - Collection of speech data
  - ASR systems today
  - ASR applications
  - ASR courses

Courses at TKK

- At Dep. Information and Computer Science:
  - **Speech recognition**
  - Natural language processing
  - Seminar courses
- At Dep. Acoustics and Signal Processing:
  - Speech processing
  - Seminar courses

Speech Recognition, 5 cr

- Goals:
  - Become familiar with speech recognition methods and applications
  - Learn the structure of a speech recognition system
  - Learn to construct one in practice!
- Period II (Nov-Dec):
  - Theory lecture every Wednesday
  - Computer lecture every Friday
  - Home works and a Project (group) work
  - No exam!

Thanks for listening…

- Contact: *Mikko.Kurimo @ tkk.fi*
- Publications, projects, demos etc:

  http://www.cis.hut.fi/projects/speech/