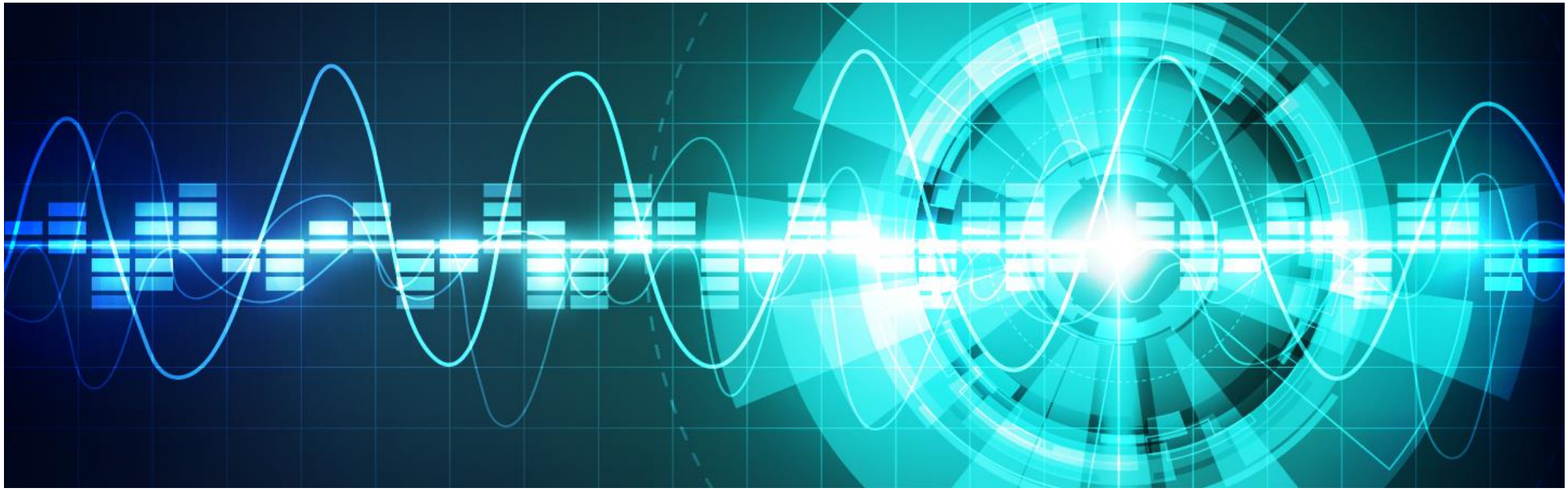
AUTOMATIC SPEECH RECOGNITION

What was the exact quote?



Fraunhofer IAIS

Creation of a Speech Manuscript

Fraunhofer IAIS

Dictate Software

internal

Fraunhofer IAIS

Search, Retrieval and Recommendation

Fraunhofer IAIS

Home Automation

Fraunhofer IAIS

Speechassistent

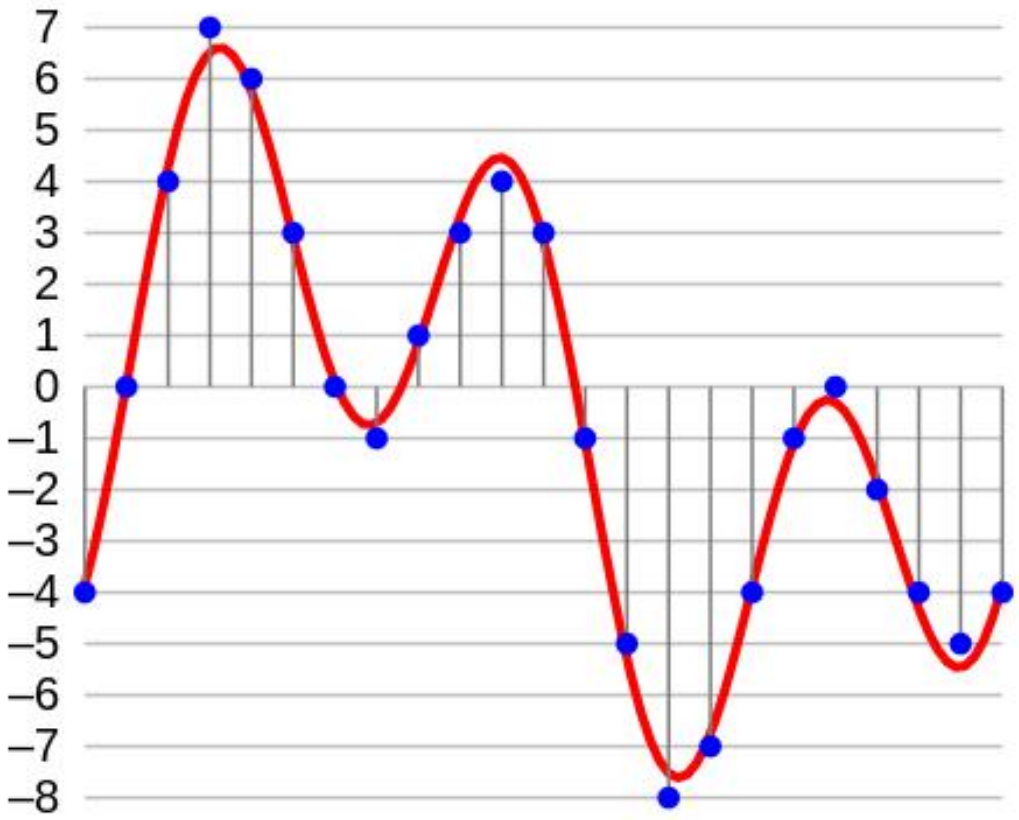
Fraunhofer IAIS

Use Cases Automatische Spracherkennung

- Creation of a speech manuscript
- Dictate software
- Subtitling of tv shows or videos
- Searchabilty of media archives (indexing)
- Home automation and speech assistents/dialog systems

- We need to differentiate between speaker dependent and independent systems. Mostly a speaker independent, continous speech recognition with a large vocabulary is required.

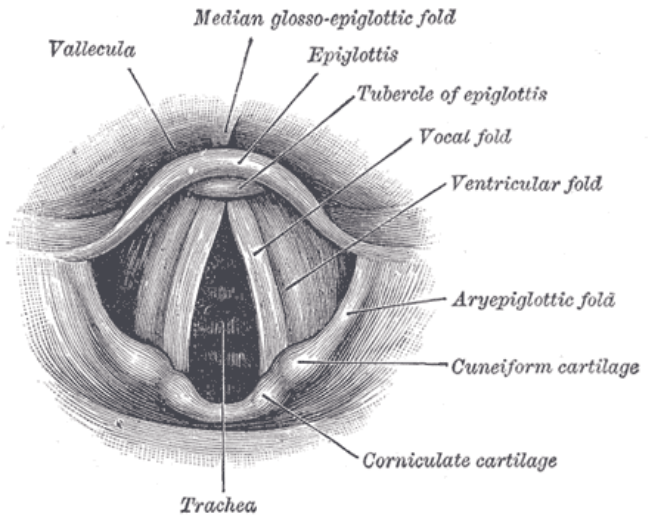
Audio Signal

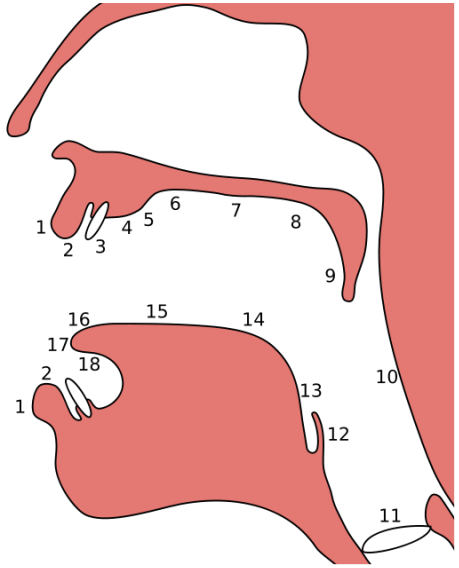


Example:
Wave-File
Pulse Code Modulation (PCM) 16bit, 16kHz

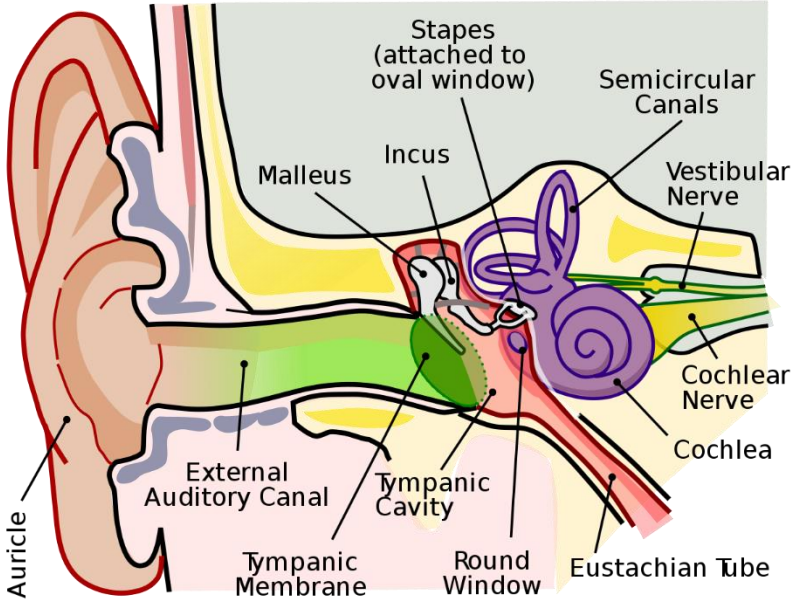
Fraunhofer IAIS

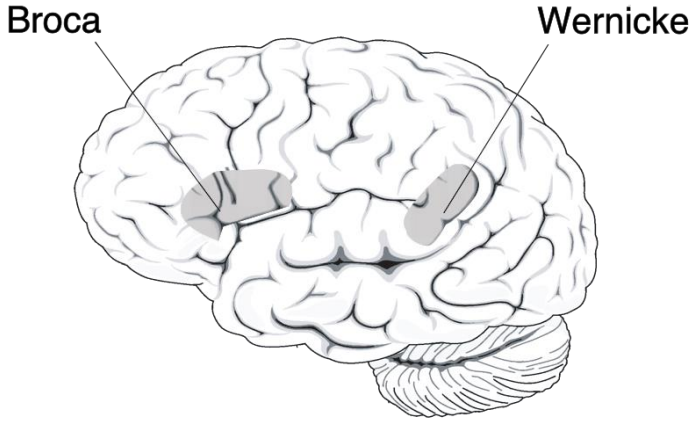
Anatomy of hearing and speaking





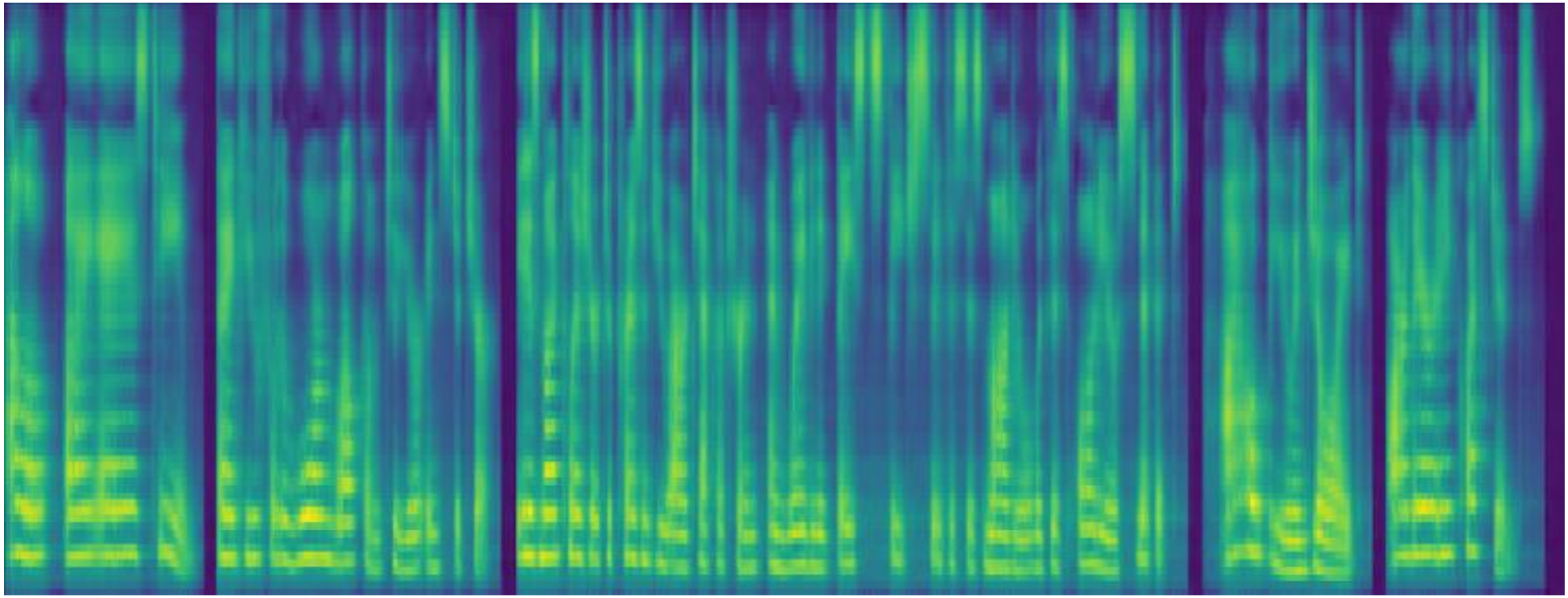
Locations of Artikulations:
1. exolabial 2. endolabial
3. dental 4. alveolar
5. postalveolar 6. präpalatal
7. palatal 8. velar 9. uvular
10. pharyngal 11. glottal
12. epiglottal 13. radikal
14. posterodorsal
15. anterodorsal 16. laminal
17. apikal 18. sublaminal





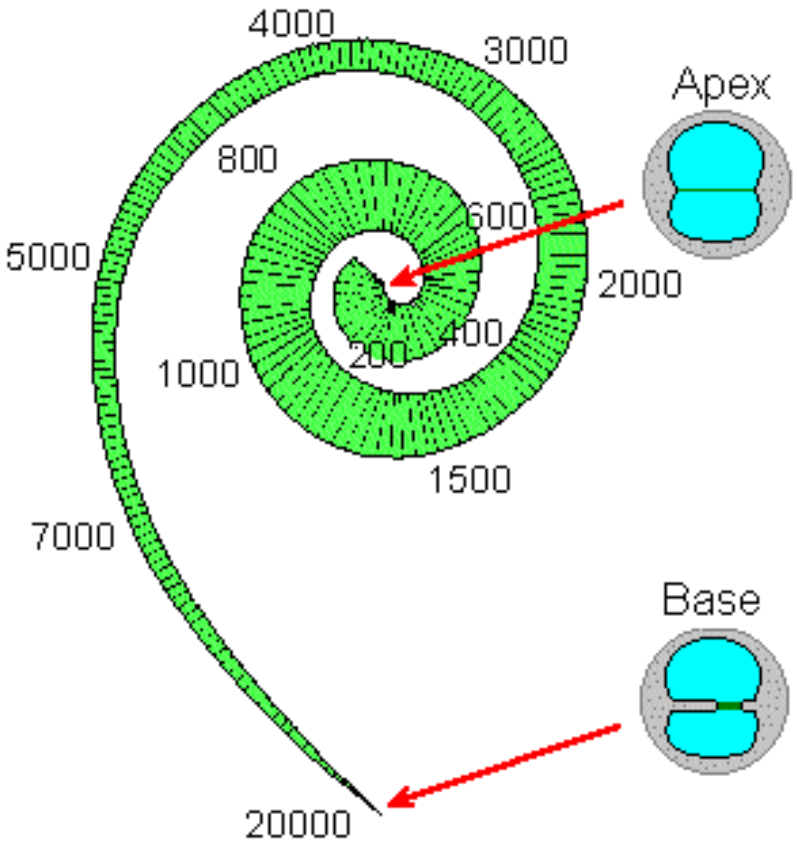
Fraunhofer IAIS

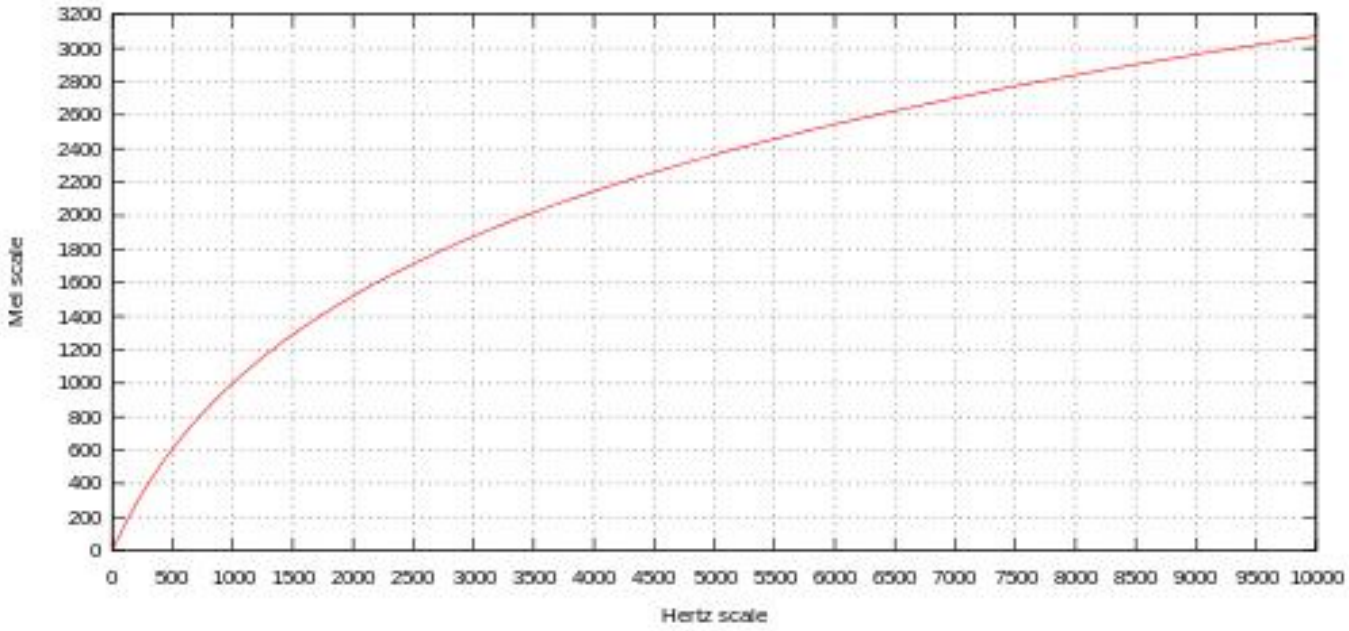
Spectrum



Fraunhofer IAIS

Acoustic Analysis – Features Mel Frequency Cepstral Coefficients (MFCCs)

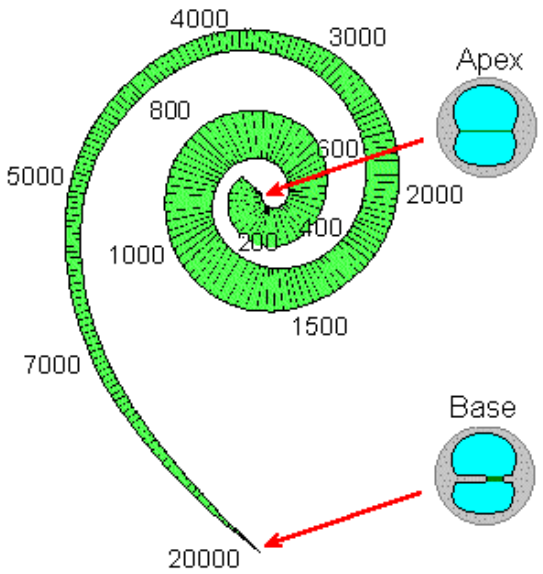


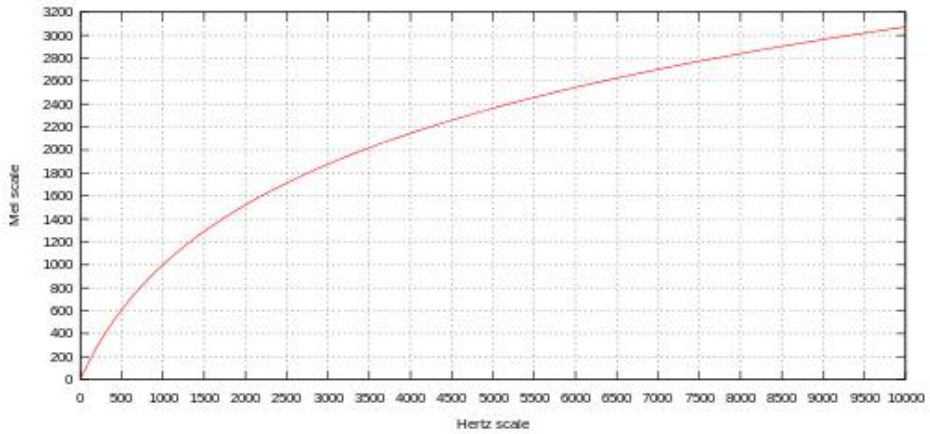


Fraunhofer IAIS

Acoustic Analysis – Features Mel Frequency Cepstral Coefficients (MFCCs)

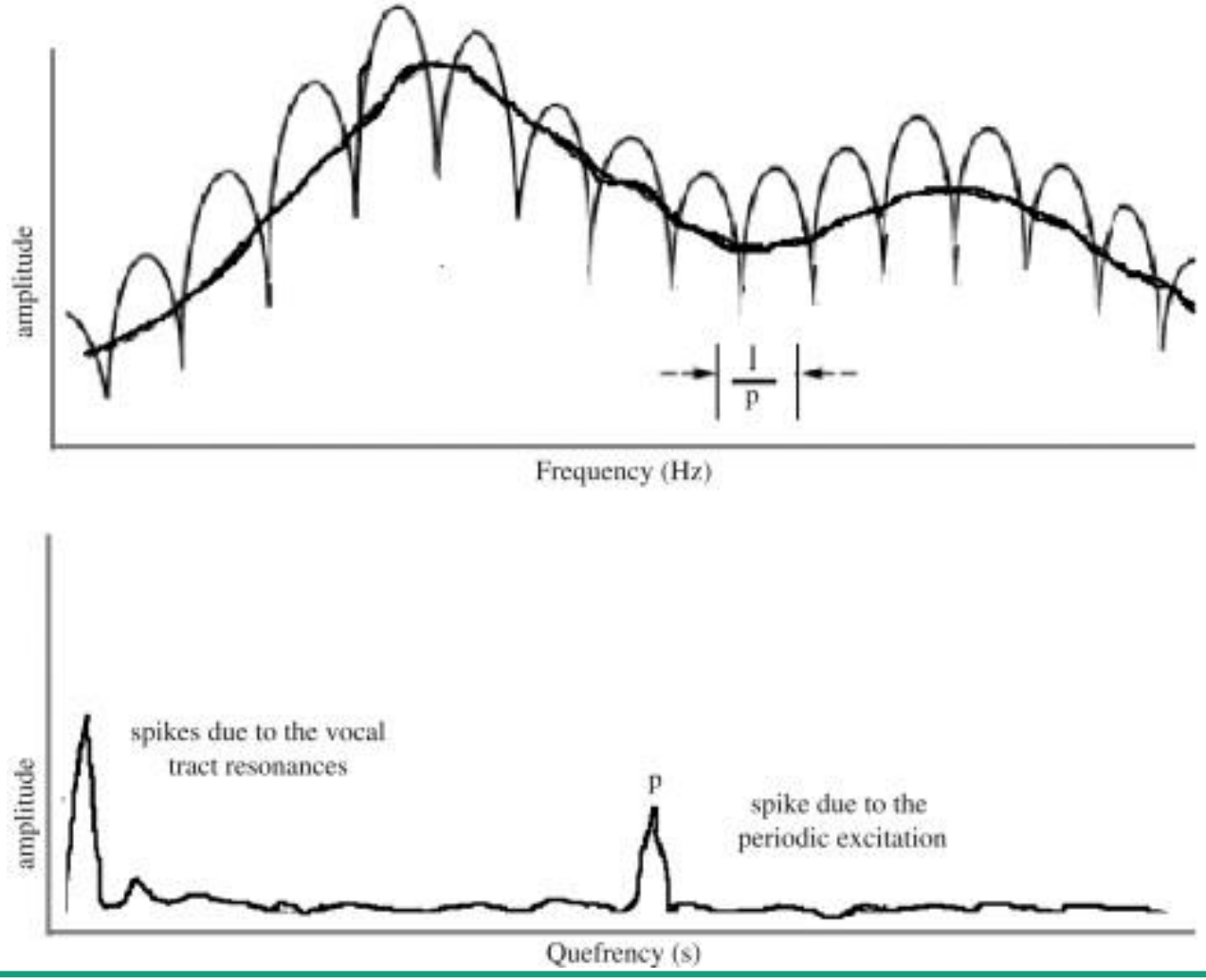
- Resolution of the human ear is different for different frequencies
- Mel-Scale considers those biological effects
- Features“:
  Mel Frequency Cepstral Coefficients (MFCCs)
  - Uses a windowed Fourier-Transformation
  - Mel-scale
    (logarithmic human perception)
  - Decorreclation of the results by using a discrete cosine transformation





Fraunhofer IAIS

Cepstrum – Spectrum of the logarithmic spectrum

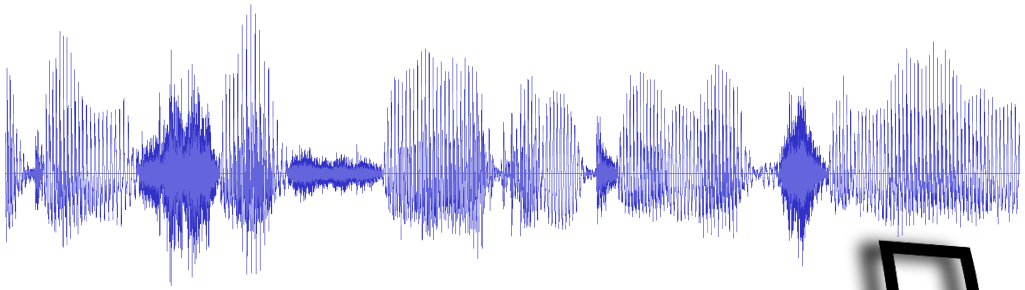


Phonemes

Sentance	The	strangers		talked		to	the	players
Phrase	The	strangers		talked		to	the	players
Word	The	strangers		talked		to	the	players
Morpheme	The	strange	er s	talk	ed	to	the	play er s
Phoneme	ðə	streynǰ	ər z	tok	t	tuw	ðə	pləy ər z

Fraunhofer IAIS

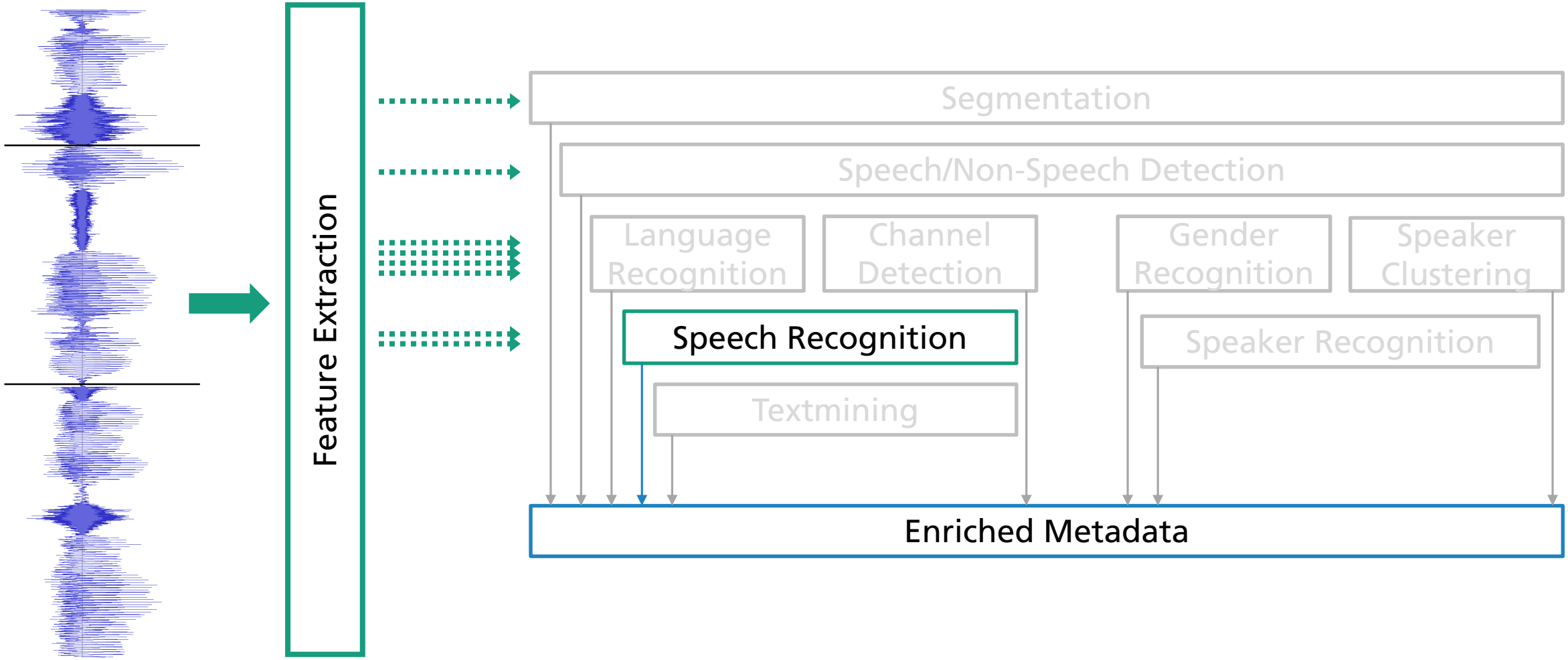
Automatic Speech Recognition – Machine Learning



Hello, my name is Jane Doe.

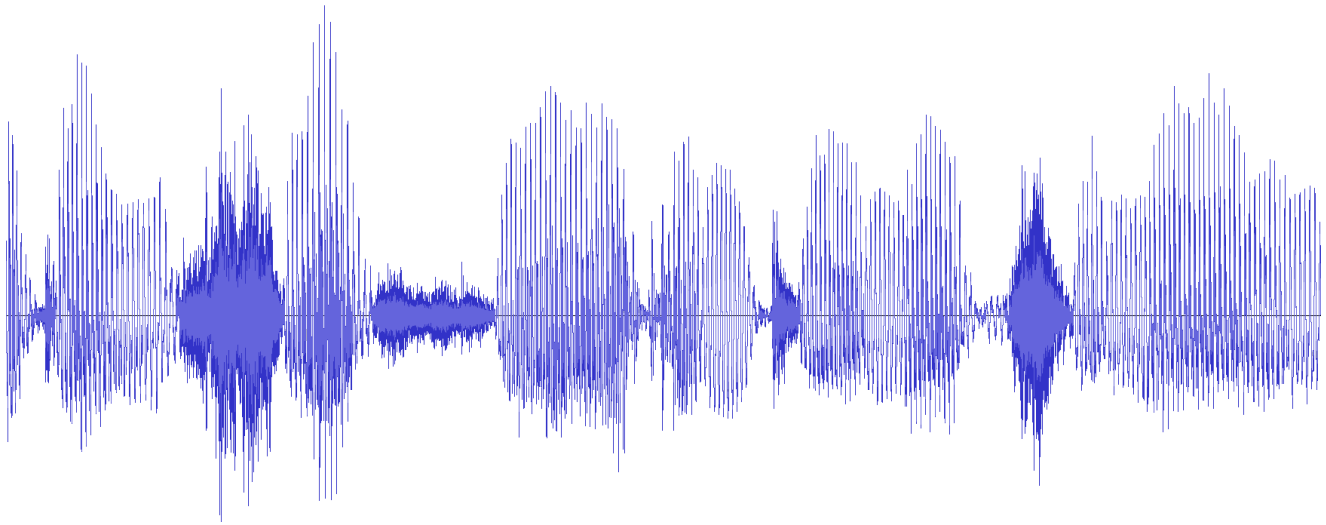
Fraunhofer IAIS

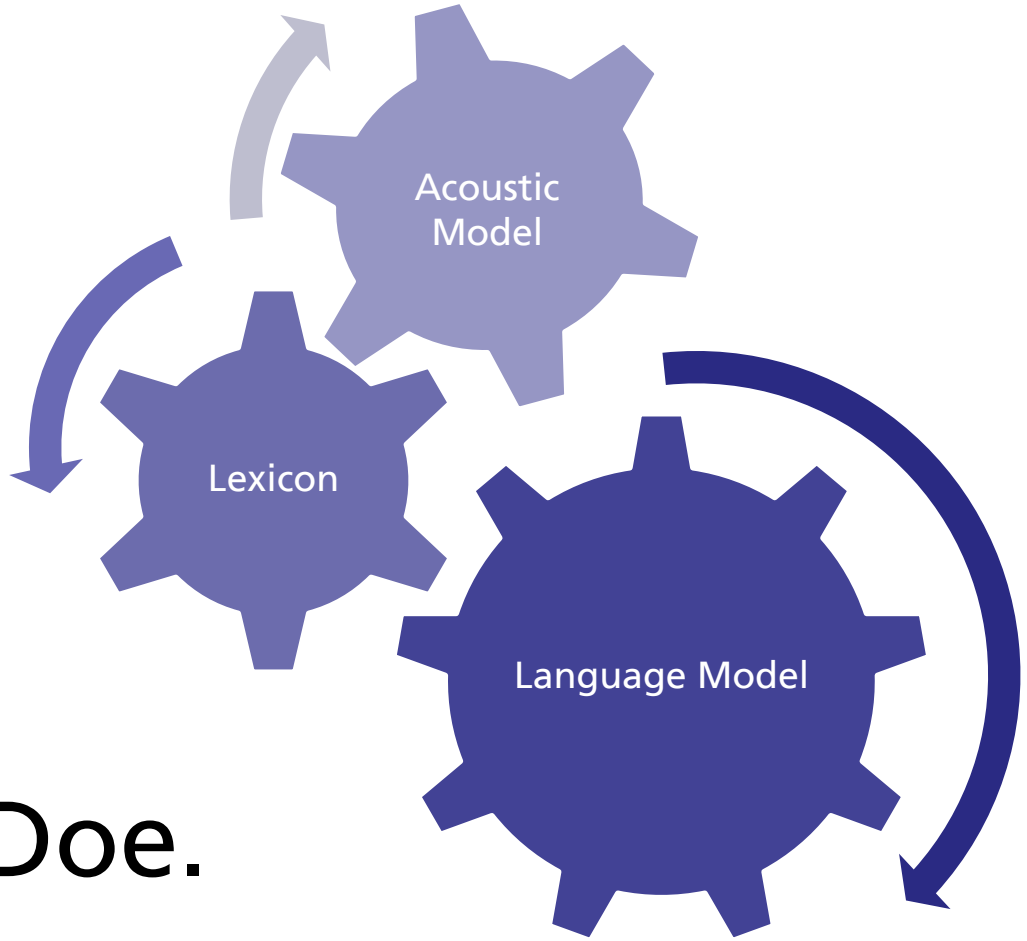
Architecture



Fraunhofer IAIS

Automatic Speech Recognition – Phoneme based statistical models





Hello, my name is Jane Doe.

Fraunhofer IAIS

Speech Recognition - Dictionary

Dictionary

/‘læŋgwidʒ/ := language

/‘læŋgwidʒiz/ := languages

...

The **dictionary** contains all words, which are to be recognized.

Filling it is done automatically.

Words which are not present in the dictionary can **not** be recognized.

Fraunhofer IAIS

Speech Recognition - Dictionary

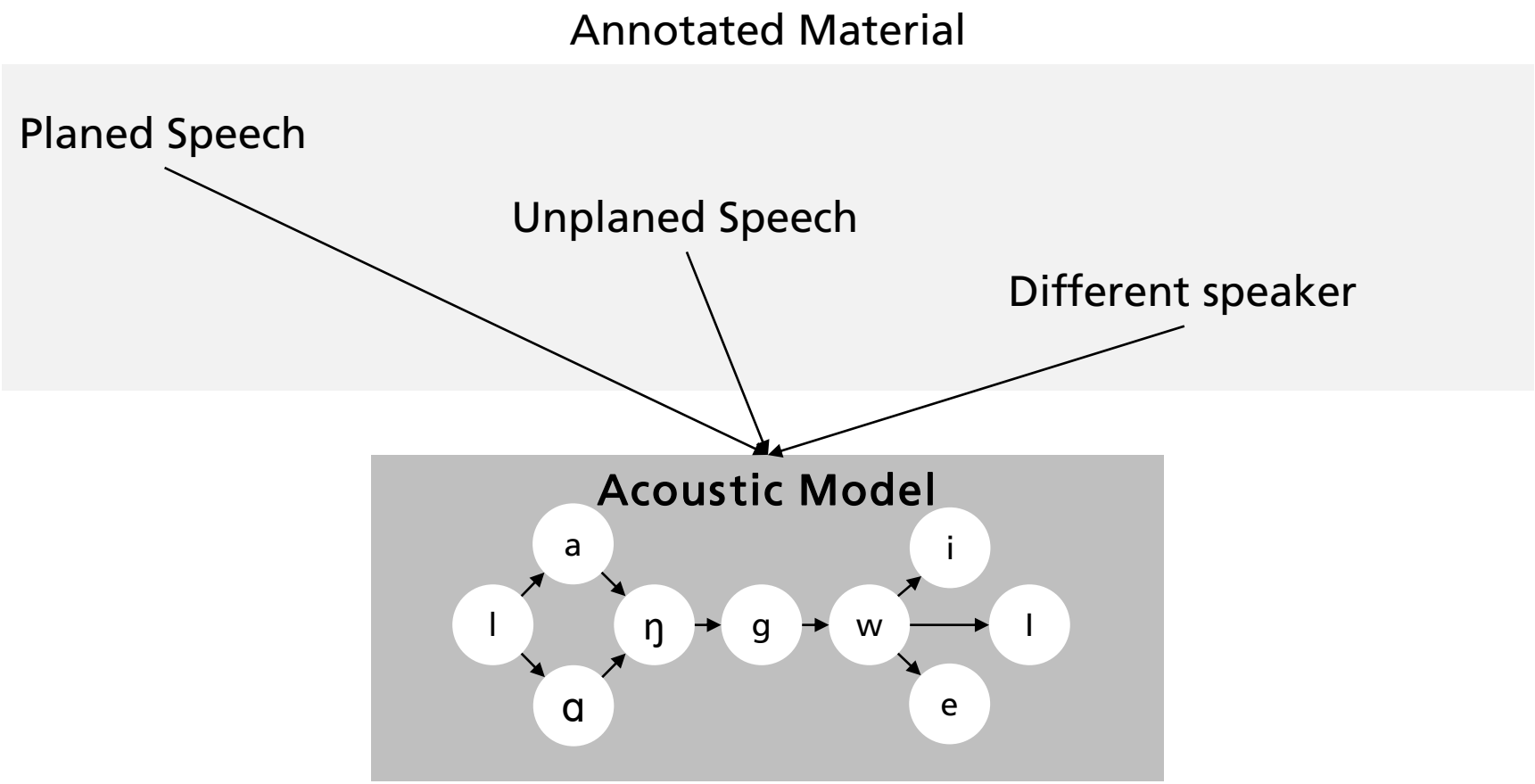
The dictionary contains all words (spelling) as well as their pronunciation. The pronunciation is partly created from manual annotations, for the most part an automatic generation of pronunciations is used. That automatic generation can easily create incorrect pronunciations for proper names.

Words which contain incorrect pronunciations in the dictionary can only be recognized reliably if they are pronounced equaly incorrect in the audio signal.

Words which are not present in the dictionary can **not** be recognized during speech recognition, as the system would not know the pronounciation of that word.

Filling the dictionary is usually done based on crawls of news pages – the same source as for the language model – to have an identical vocabular as the language model. However, this may lead to errors as typos are added to the dictionary as well – if the same mistake was done several times.

Speech Recognition – Acoustic Model



Translation of the Audio Signal into chains of phonems.

Fraunhofer IAIS

Speech Recognition – Acoustic Model

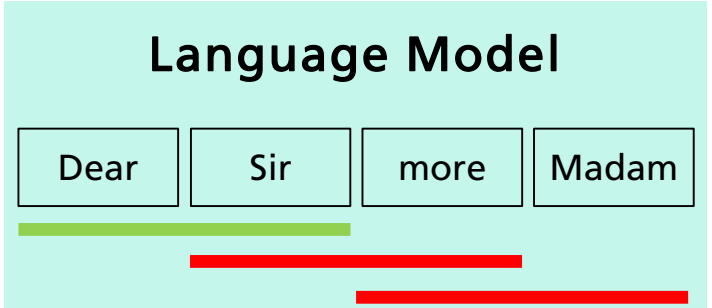
The acoustic model is used for the „translation“ of the speech signal into phonemes.

Typically this translation is nowadays done by deep neural networks. The training of the networks requires large amount of annotated audio material containing speech– at least 100h, better 1000+h.

The speech material should contain different recording situations as:

- Clean recordings:
  - Trained speaker
  - No (or low levels of) noise
  - Good modulation
- Unclean recordings:
  - Untrained speaker, unclean pronunciation
  - Different kinds and intensity of noice
  - Different kinds of disturbances like background noice, music, jingle, ..

Speech Recognition – Language Model



All words which are not present in the language model can **not** be recognized.

Data for the **Language Model** are gathered by crawling news pages.

Fraunhofer IAIS

Speech Recognition – Language Model

The modeling of the probability of different word sequences is done by the language model.

Trainings data for this model is usually gathered by crawling news pages. This creates a good representation for planed speech, as it is very similar. However this results in less optimal recognitions for spontaneous speech, where the speaker might repeat or omit words or end sentances prematurely.

The recognition of unseen word sequences is panalized by the language model, as those word sequences are treated as unlikely.

As the used language differs between domains it is reasonable to use different language models for different domains (e.g. news, sport, talkshows, ..).

Speech Recognition – Challanges

Hardware

Room Acoustics

Multiple Speaker

Challanges

Background Noise

Paralinguistic

Spontaneous Speech

Fraunhofer IAIS

Speech Recognition - Challanges

Hardware

- Poor recording quality (microphon)
- Introducing artifacts by encoding the signal

Room acoustic

- Especially reveberation lead to degrated quality of the speech recognition. For the recognized it „sounds“ as if two speakers are speaking at the same time.

Multiple Speaker

- The signals of multiple speakers are often not separable, so that the speech recognizer can’t work on individual speech signals.

Speech Recognition - Challanges

Background noises

- Music, jingle, street noise, talk in the background, .. lead to degradation of the speech recognition quality, in case that the acoustic model did not see similar data in the training corpus.

Paralinguastic

- Especially whispering and shouting sound very different from normal speech prodution. However those special situations are rarely contained in the trainings material, making it hard to recognize such speech.

Spontaneous speech

- Unclean pronunciation, different sentence construction and unusual vocabulary are characteristic for spontaneous speech. To improve the recognition rates for such situations, the trainings data needs to contains comparable data.

Benchmarking

internal

Fraunhofer IAIS

Benchmarking

- To benchmark speech recognition results you require a suitable testset, i.e. a testset which is close to your real application
- Ensure a correct annotation of your testset, meaning you decide how you want to annotate phrases like „I'll do that" versus „I will do that"

Error Rates

Word Error Rate: 25%

Judge Brett M. Kavanaugh faced a whirlwind of new accusations on Wednesday.												(annotation)
Judge Brad m Kavanaugh face a whirlwind _ new accusations on Wednesday												(hypothesis1)
M	S	S	M	S	M	M	D	M	M	M	M	

Word Error Rate: 12.5%

Judge Brad m Kavanaugh faced a whirlwind of new accusations on Wednesday												(hypothesis2)
M	S	S	M	M	M	M	M	M	M	M	M	

Entity Error Rate: 100%

Judge Brett M. Kavanaugh faced a whirlwind of new accusations on Wednesday.												(annotation)
Judge Brad m Kavanaugh face a whirlwind new accusations on Wednesday												(hypothesis1)
		S										

Fraunhofer IAIS

Open Source Software

Fraunhofer IAIS

Hinweise zur Open Source Software

- Sphinx
  - Speech recognition systems developed at Carnegie Mellon University (pre-trained data included)
- HTK (Hidden Markov Model Toolkit)
  - HMM training for speech recognition
- Julius
  - Stable, commercial decoder (without language model and acoustic model)
- KALDI
  - State-of-the-art toolkit (including Deep Neural Networks) for automatic speech recognition
  - Quellcode in C++, verfügbar unter der Apache Lizenz

Who am I?

- **David Laqua**
- Research Engineer
- Telefon: +49 2241 / 14 2725
- Email: david.laqua@iais.fraunhofer.de

internal

Fraunhofer IAIS

Disclaimer

Responsible contact: David Laqua
E-mail: david.laqua@iais.fraunhofer.de