technische universität dortmund
Institut für Sprache, Literatur und Kultur

Seeking patterns in language: Using ASR technology for linguistic studies

TCLC University of Trier | 14 May 2021

Andreas Weilinghoff

Table of Contents

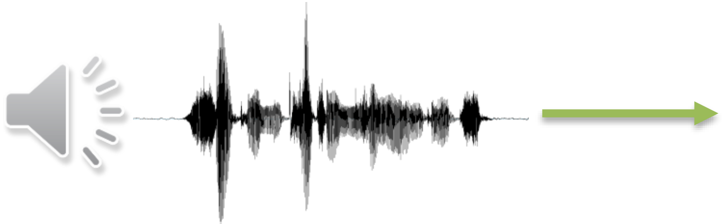
01 INTRO	• What is ASR? • How does ASR work?
02 EXAMPLES	• ASR Services • Speeding up transcription work
03 DISCUSSION	• Advantages • Limitations
04 REFERENCES	• Bibliography • Summary of useful resources

01 INTRO

What is ASR?

“Automatic speech recognition (ASR) is the process and the related technology for converting the speech signal into its corresponding sequence of words or other linguistic entities by means of algorithms implemented in a device, a computer, or computer clusters.”

(Deng and O’Shaughnessy 2003; Huang et al. 2001 cited in Li et al. 2016)



`seeking patterns in language`

- research field for roughly 70 years

What is ASR?

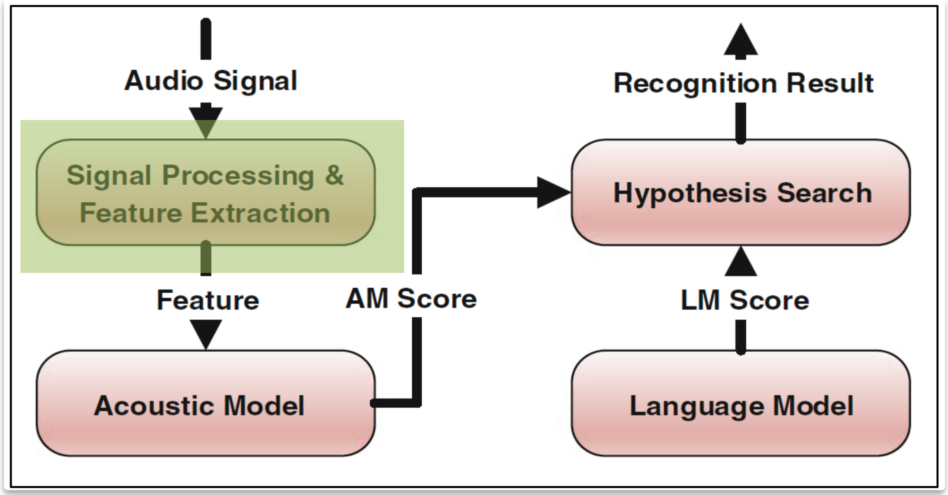
- great advancements in recent years due to

→ exponential growth of data
→ drastic increase of computing power
→ successful implementation of neural networks

applications: voice search, personal digital assistance systems (PDA), automated captioning, gaming etc.

transcription work ?!

How does ASR work?

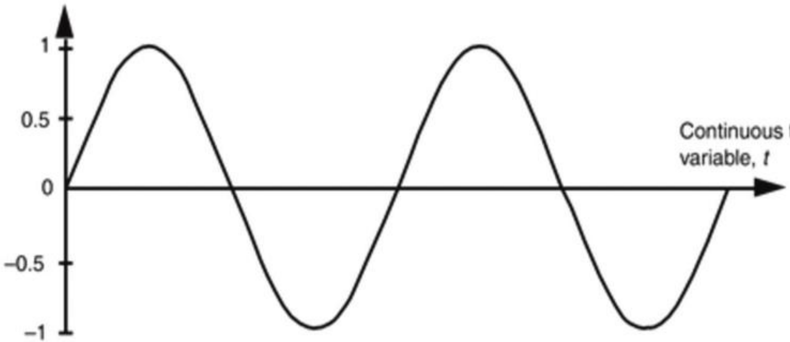


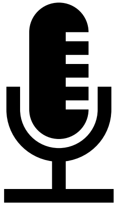
(Yu and Deng 2015: 4)

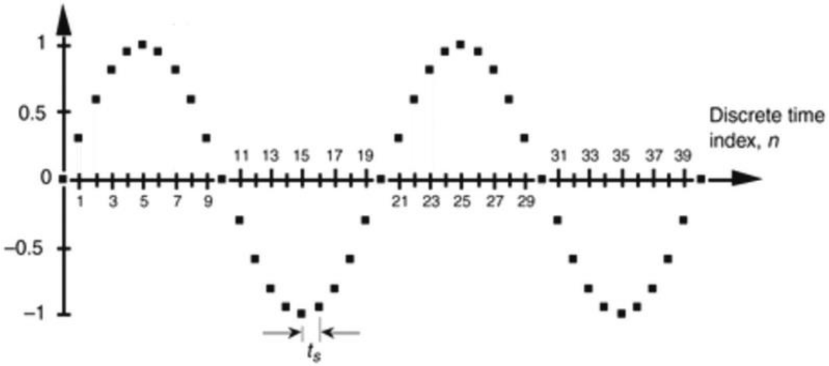
Basic architecture of ASR systems

How does ASR work?

Step 1: From analogue to digital





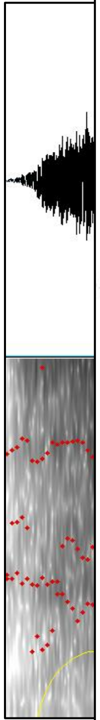


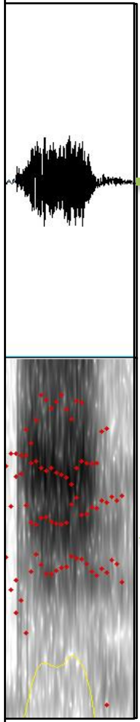
continuous (analog) air pressure variations (soundwaves)

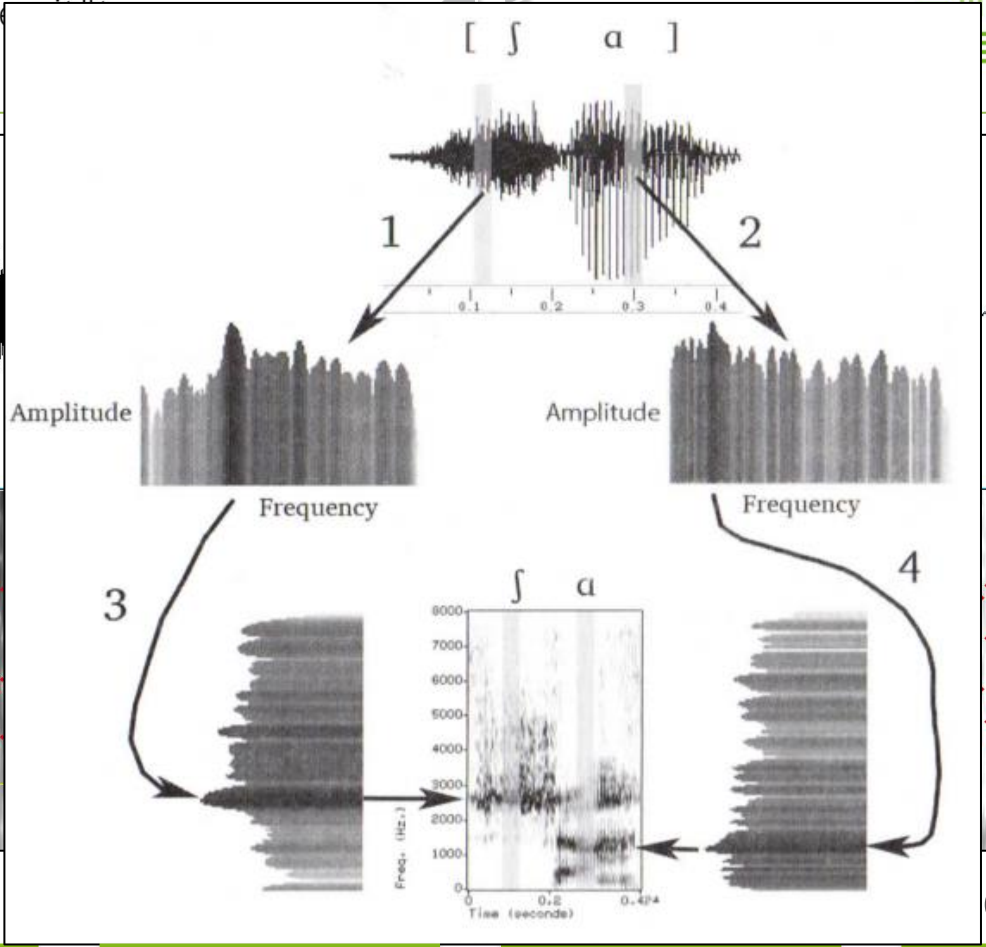
discrete (digital) representation of soundwave in a particular sampling frequency

signal
input
(waveform)

broadband
spectogram

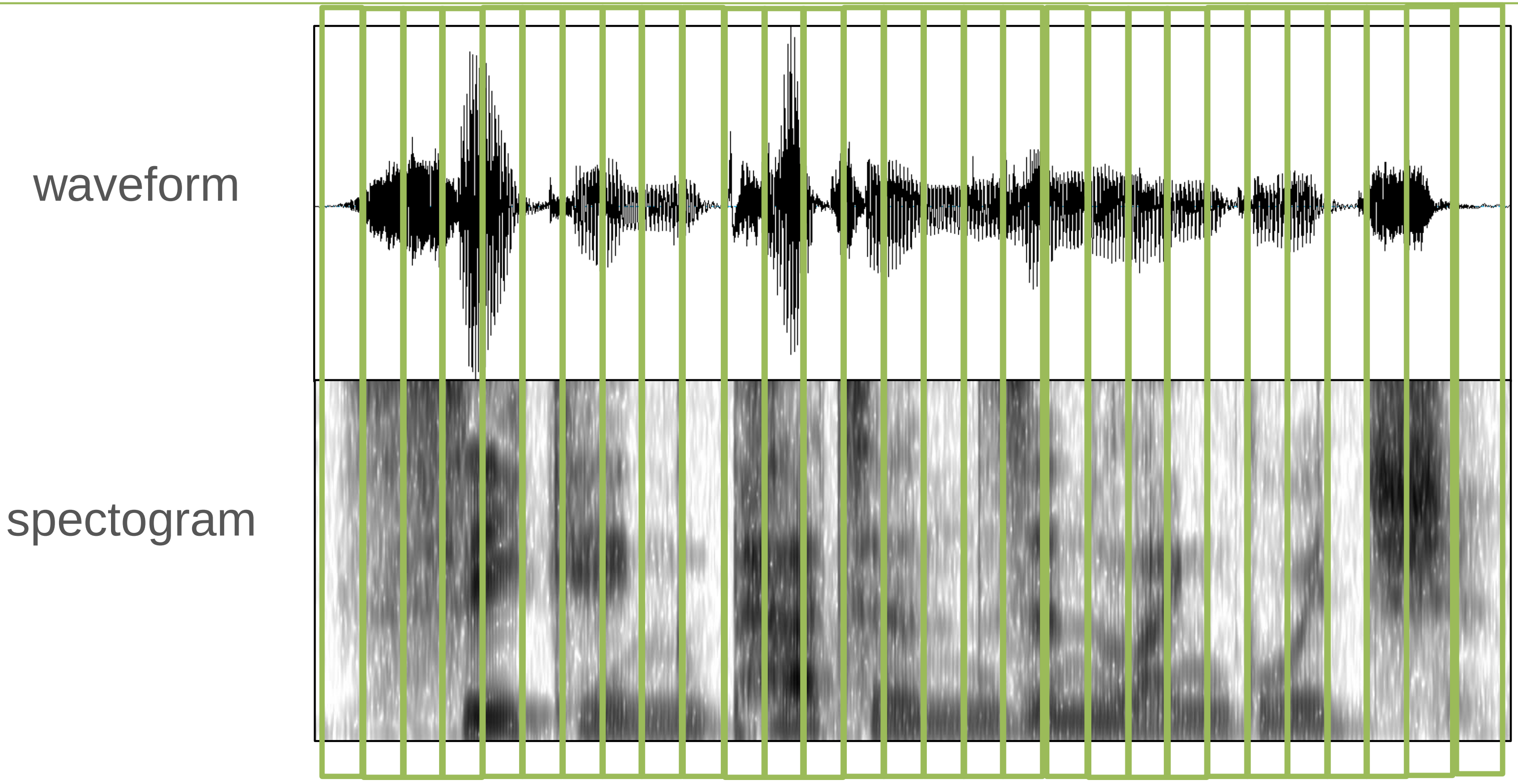




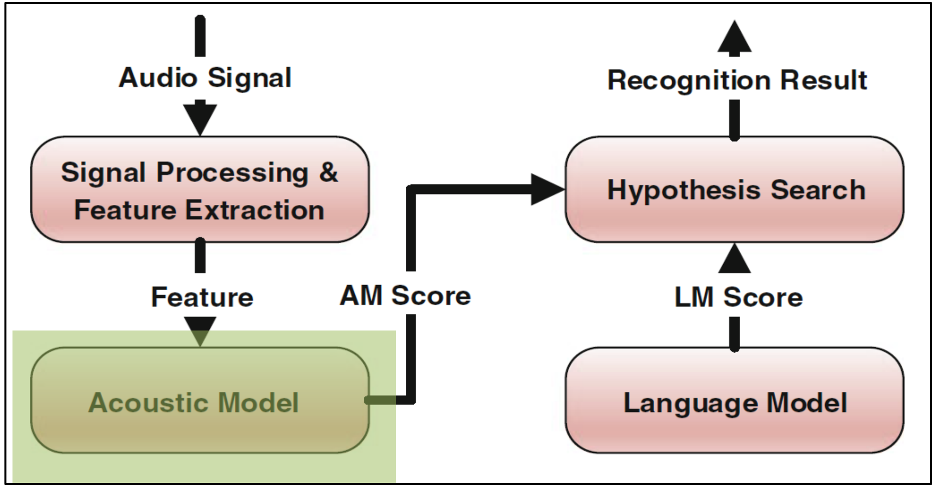


Fourier transformation

(Ladefoged and Johnson 2015: 9)


waveform
spectogram

How does ASR work?



(Yu and Deng 2015: 4)

How does ASR work?

Step 2: Acoustic model

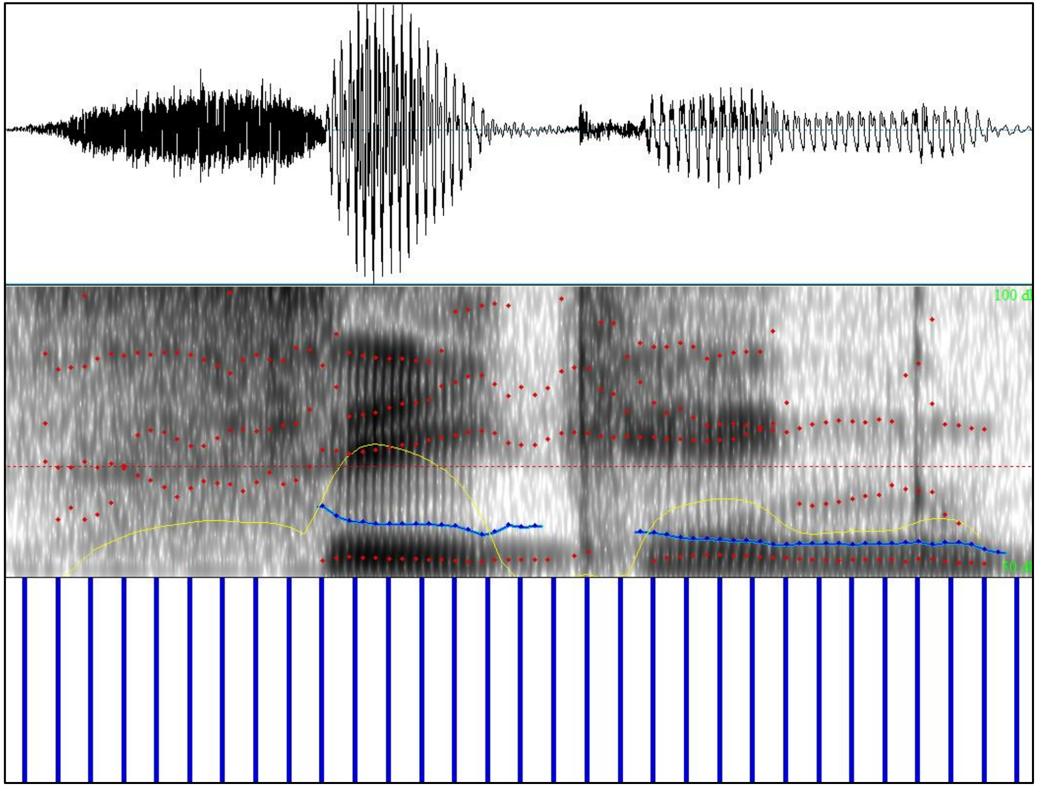
Hidden Markov Models (HMM)

Deep Neural Networks (DNN)

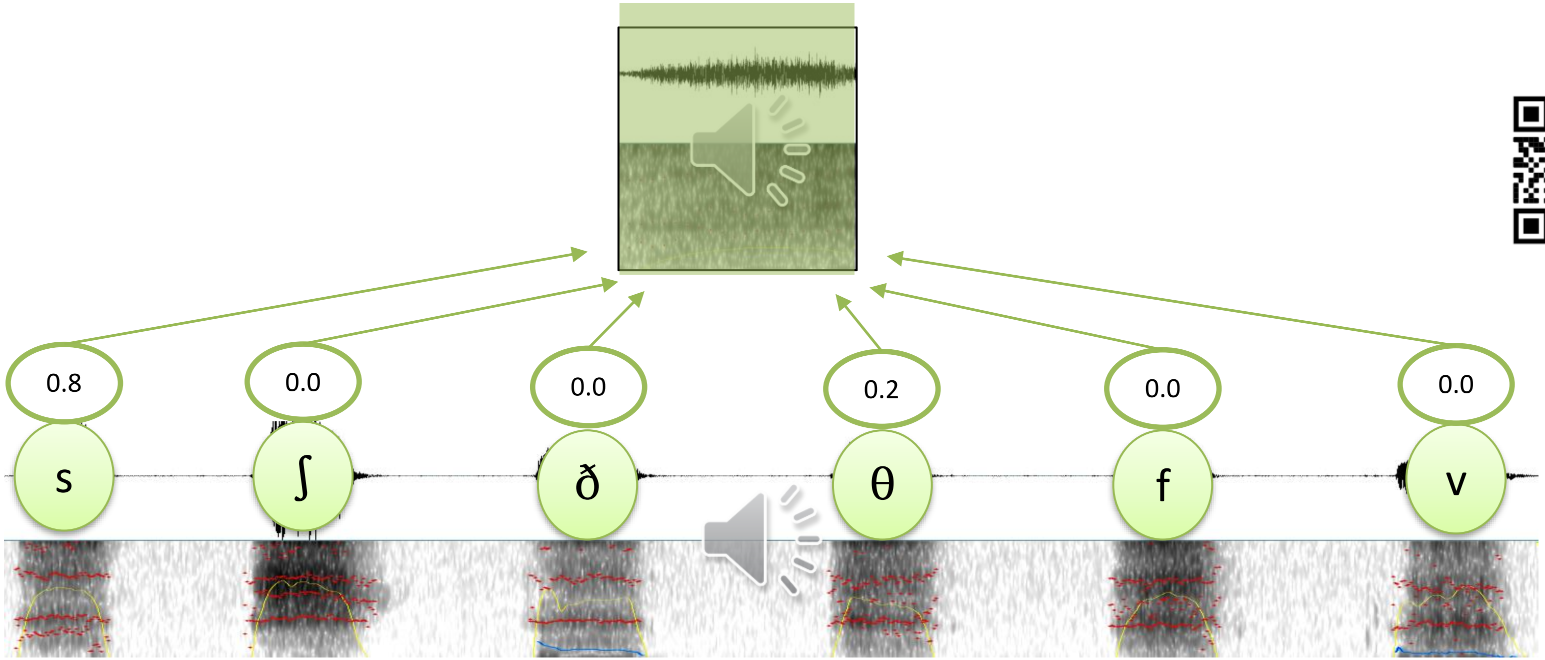
Phonetic Reference Dictionary (Pronunciation Model)

What is the most likely phone/phoneme given the processed audio input in a particular timeframe?

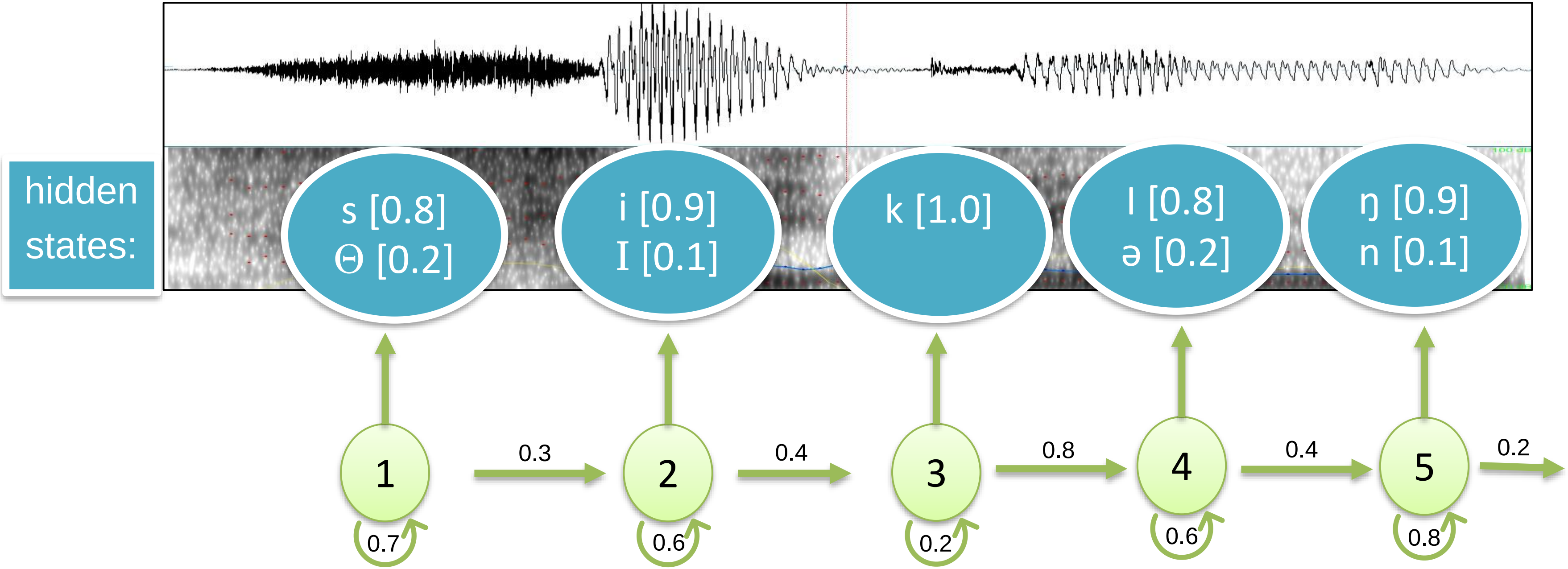
Sampled audio input



Hidden Markov Models (HMM)

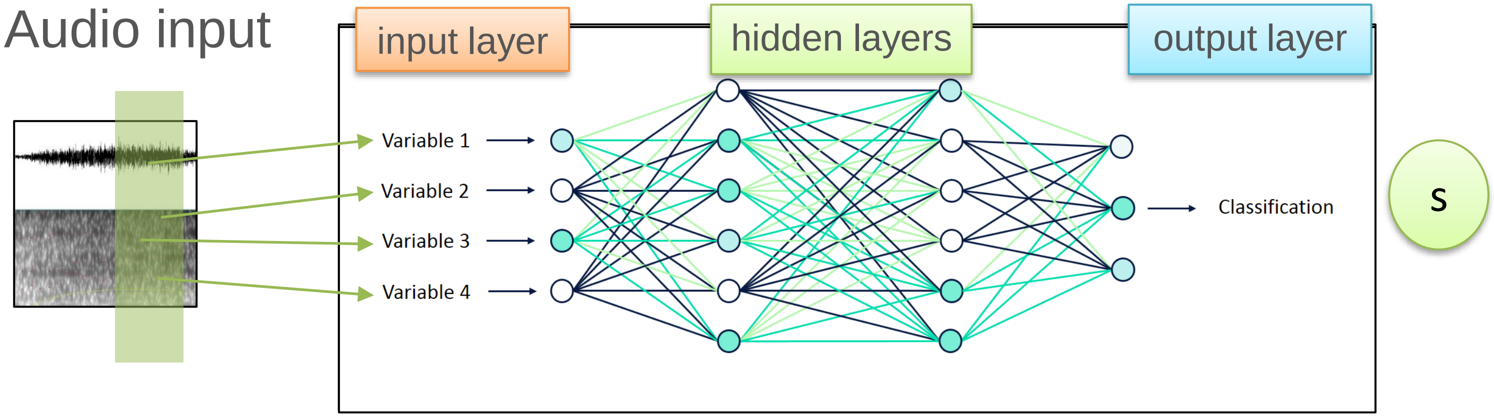


Hidden Markov Models (HMM)



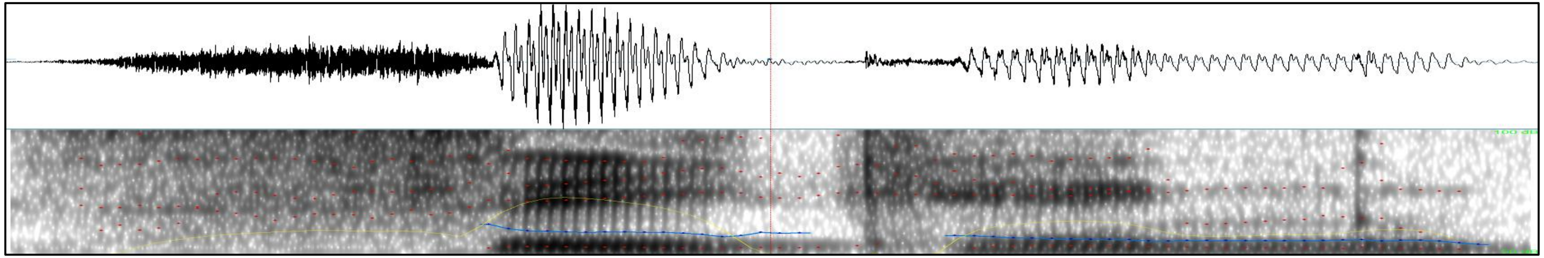
Deep Neural Networks (DNN)

Audio input



Retrieved from: https://community.alteryx.com/t5/Data-Science/It-s-a-No-Brainer-An-Introduction-to-Neural-Networks/ba-p/300479

How does ASR work?



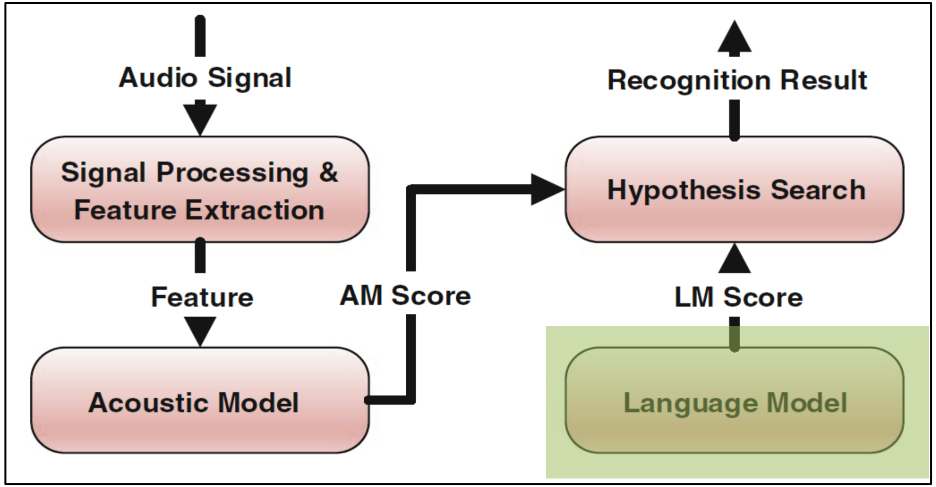
HMM/DNN: S IY K IH NG

Phonetic Reference Dictionary (Pronunciation Model)

S IY K IH NG

sea king
C. King
seeking
see king

How does ASR work?



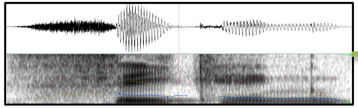
(Yu and Deng 2015: 4)

Language Model

Acoustic model

Language model

Audio input



S IY K IH NG

sea king

C. King

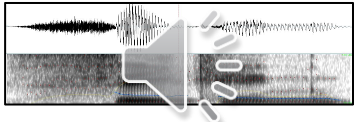
seeking

see king

How likely is it for different words to occur (together)?

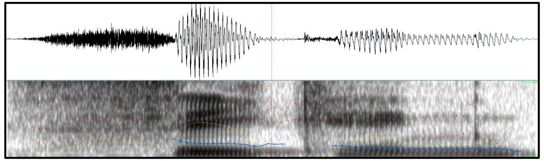
Language Model

Audio input

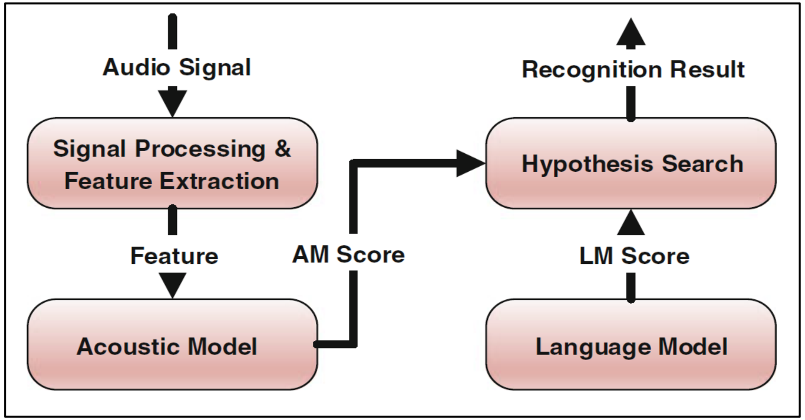


Acoustic model	Language model
AH D UW DH IH S EH M V ER IY W IY K	I do this uh very weak.
AY D UW DH IH S EH V ER IY W IY K	I do this $\boxed{\text{every week.}}$
T UW IH S V ER IY W IY K	Two is very weak.

How does ASR work?



seeking
patterns in
language



(Yu and Deng 2015: 4)

02 EXAMPLES

ASR Services

(Partly) Commercial:

→ IBM Watson STT

→ Microsoft Azure STT

→ Google Cloud STT

→ Amazon AWS STT

...

ASR Services

Non-commercial:

- → HTK Toolkit (University of Cambridge)
- → Carnegie Mellon University Sphinx toolkit
- → Kaldi toolkit

Further platforms/libraries: Common Voice (Mozilla), Tensorflow (Google)

IBM Watson STT

- state of the art ASR service
- long soundfile transcription possible
- includes timestamps and word alignment
- robustness & different language models (e.g AusE, BrE, AmE)
- user-friendly, adaptable and well documented
- free of charge (500 minutes per month for Lite users)

→ Accessible for academic users via WebMAUS interface

Speeding up transcription work

Usual approach:

1. Collecting data
2. Broad transcriptions
3. Further analyses and preparation (e.g. narrow transcriptions)

Advantages: precise and flexible human analysis

Disadvantages: time-consuming, work intensive, tedious, human errors

(~ 6 min of sound file = 1 h transcription work)

Speeding up transcription work

- Apply ASR for broad transcriptions
- Forced alignment and automatic syllabification parsing
- …

03 DISCUSSION

Advantages

→ automatic pre-processing of data

→ time saving

→ assistance in transcription work

→ less “random“ errors → more consistent

Limitations

... the higher the audio quality

... the more structured the speech

... the more ‘standard‘ the speech

... the less speakers involved?

... the better

Limitations

- → depending on the research project, manual checking and correcting remains necessary
- → data protection?

Discussion

How could AI affect our research areas?

Discussion

Could the role of the researcher change?

04 REFERENCES

References

Boersma, Paul & Weenink, David (2021). *Praat: doing phonetics by computer* [Computer program]. Version 6.1.42, retrieved 15 April 2021 from http://www.praat.org/

Deng, L., O'Shaughnessy, D. (2003). *Speech Processing – A Dynamic and Optimization-Oriented Approach.* CRC Press.

ELAN (Version 6.0) [Computer software]. (2020). Nijmegen: Max Planck Institute for Psycholinguistics, The Language Archive. Retrieved from https://archive.mpi.nl/tla/elan

HTK (Version 3.4.1) [Software]. (2009). University of Cambridge. Retrieved from http://htk.eng.cam.ac.uk

Huang, X., Acero, A., Hon, H.W. (2001). *Spoken Language Processing: A guide to theory, algorithm, and system development.* Prentice hall PTR.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G., Vesely, K. (2011). *The Kaldi Speech Recognition Toolkit*. IEE Signal Processing Society.

References

Kisler, T., Reichel U. D. & Schiel, F. (2017): Multilingual processing of speech via web services, Computer Speech & Language, 45, 326–347.

Ladefoged, P., Johnson. K. (2015). A Course in Phonetics. Stamford: Cengage Publishing.

Lamere, P., Kwok, P., Gouv, E. B., Singh, R., Walker, W., Wolf, P. (2003). *The CMU SPHINX-4 speech recognition system*, IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP 2003), Hong Kong, 1, pp. 2–5.

Li, J., Deng, L., Haeb-Umbach, R., Gong, Y. (2016). *Robust Automatic Speech Recognition: A Bridge to Practical Applications.* Elsevier Publishing.

Microsoft Research. (2017). Automatic Speech Recognition: An Overview. [Video]. Youtube. https://www.youtube.com/watch?v=q67z7PTGRi8

Yu, D., Deng, L. (2015). *Automatic Speech Recognition: A Deep Learning Approach*. London: Springer Publishing.

Summary of useful resources

Speech processing / forced alignment:

BAS Webservices (WebMAUS) https://clarin.phonetik.uni-muenchen.de/BASWebServices/

DARLA http://darla.dartmouth.edu/cave

Montreal Forced Aligner https://github.com/MontrealCorpusTools/Montreal-Forced-Aligner

FAVE Aligner https://github.com/JoFrhwld/FAVE

Summary of useful resources

Toolkits:

HTK https://htk.eng.cam.ac.uk

CMUSphinx https://cmusphinx.github.io/

Kaldi ASR https://kaldi-asr.org/

Summary of useful resources

IBM Watson:

IBM Watson STT

https://www.ibm.com/cloud/watson-speech-to-text

Nicolas Renotte

https://www.nicholasrenotte.com/

https://github.com/nicknochnack

https://www.youtube.com/channel/UCHXa4OpASJEwrHrLeIzw7Yg