Methods and Algorithms of Automatic Speech Recognition

Ronald Lencevičius

September 2021

California State Polytechnic University, Pomona

Introduction

Introduction

- Software Engineer Internship at NUANCE

Introduction

- Software Engineer Internship at NUANCE
- Worked on an internal speech recognizer (supervised by David Nolden, PhD)

Introduction

- Software Engineer Internship at NUANCE
- Worked on an internal speech recognizer (supervised by David Nolden, PhD)
    - Cleaned up code

Introduction

- Software Engineer Internship at NUANCE
- Worked on an internal speech recognizer (supervised by David Nolden, PhD)
  - Cleaned up code
  - Tested parts of decoding algorithm – on small and large scale

Introduction

- Software Engineer Internship at NUANCE
- Worked on an internal speech recognizer (supervised by David Nolden, PhD)
  - Cleaned up code
  - Tested parts of decoding algorithm – on small and large scale
  - Implemented some new additions to the algorithm

Introduction

- Software Engineer Internship at NUANCE
- Worked on an internal speech recognizer (supervised by David Nolden, PhD)
  - Cleaned up code
  - Tested parts of decoding algorithm – on small and large scale
  - Implemented some new additions to the algorithm
- Learned a ton about Speech Recognition (SR)

Introduction

Speech Recognition (SR)

- Methods
  - Statistical SR
  - End-to-End (E2E) SR
- Algorithms
  - Processing
  - Decoding

Statistical SR[1]

Statistical SR[1]

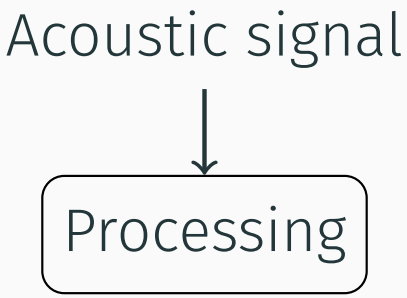
The overall goal of SR is to apply Bayes' Theorem

$$\begin{aligned}
\text{optimal words} &= \underset{words}{\arg\max}\{p(\text{words given speech})\} \\
&= \underset{words}{\arg\max}\{p(\text{speech given words}) \cdot p(\text{words})\} \\
w_{opt} &= \underset{w}{\arg\max}\{p(x|w) \cdot p(w)\}
\end{aligned}$$

Statistical SR[1]

Acoustic signal

Statistical SR[1]



Statistical SR[1] – Processing

Some comments...

Statistical SR[1] – Processing

Some comments...

- Want to minimize irrelevant information (outside noise)

Statistical SR[1] – Processing

Some comments...

- Want to minimize irrelevant information (outside noise)
  - Fast Fourier Transform

Statistical SR[1] – Processing

Some comments...

- Want to minimize irrelevant information (outside noise)
  - Fast Fourier Transform
- Normalize the transformed data

Statistical SR[1] – Processing

Some comments...

- Want to minimize irrelevant information (outside noise)
  - Fast Fourier Transform
- Normalize the transformed data
  - Pitch

Statistical SR[1] – Processing

Some comments...

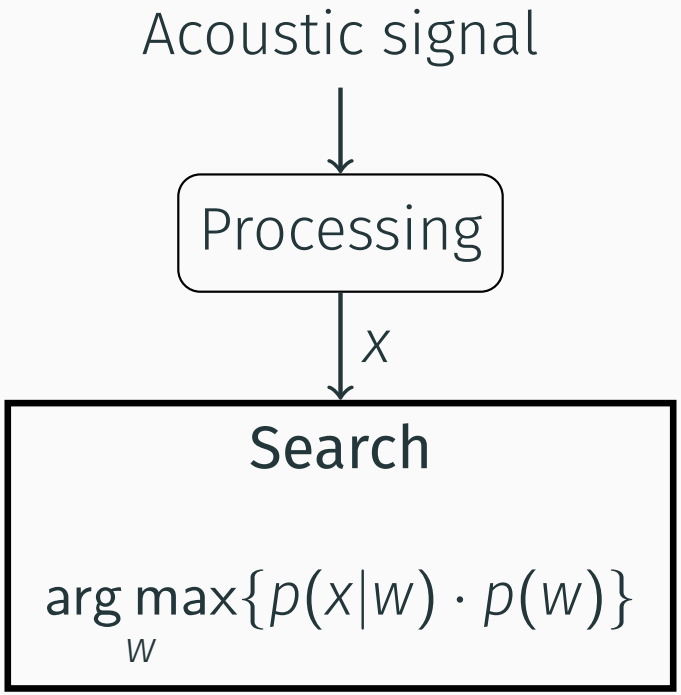
- Want to minimize irrelevant information (outside noise)
  - Fast Fourier Transform
- Normalize the transformed data
  - Pitch
  - Tone

Statistical SR[1] – Processing

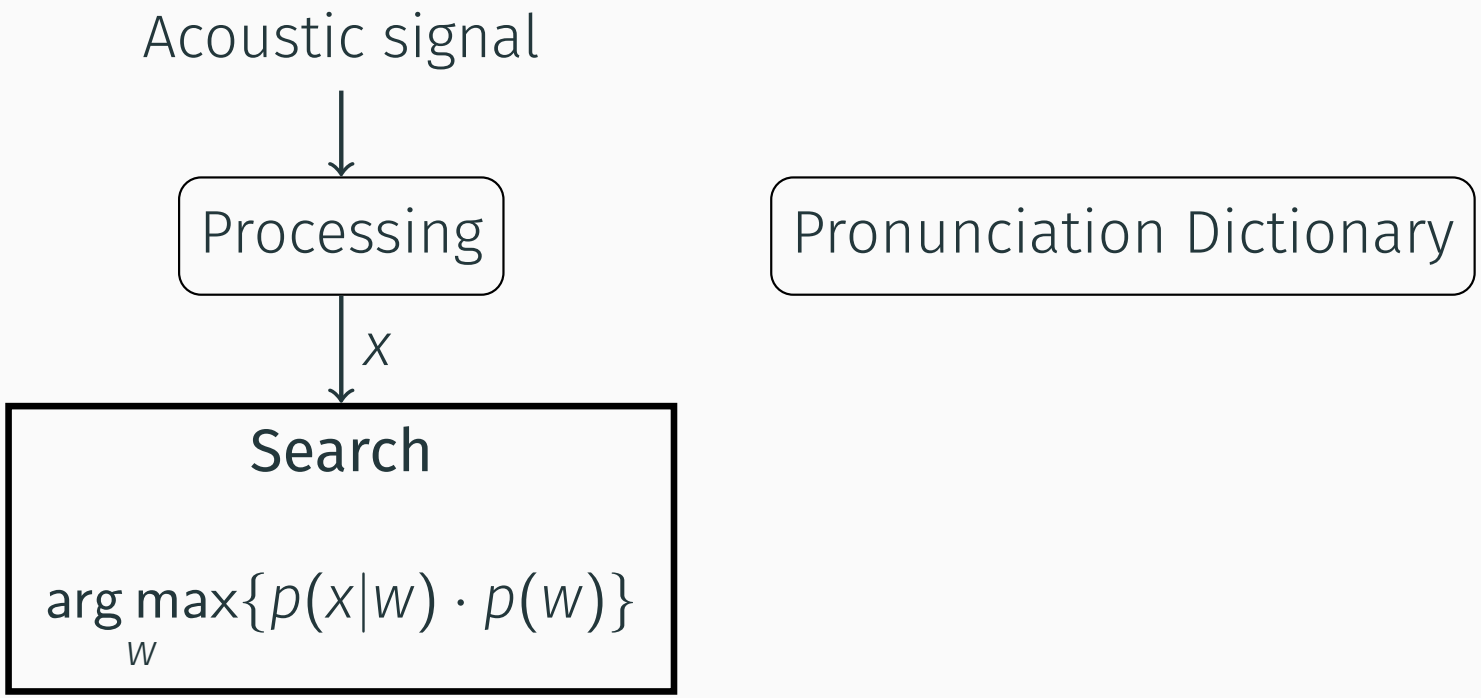
Some comments...

- Want to minimize irrelevant information (outside noise)
  - Fast Fourier Transform
- Normalize the transformed data
  - Pitch
  - Tone
  - Vocal Tract Length

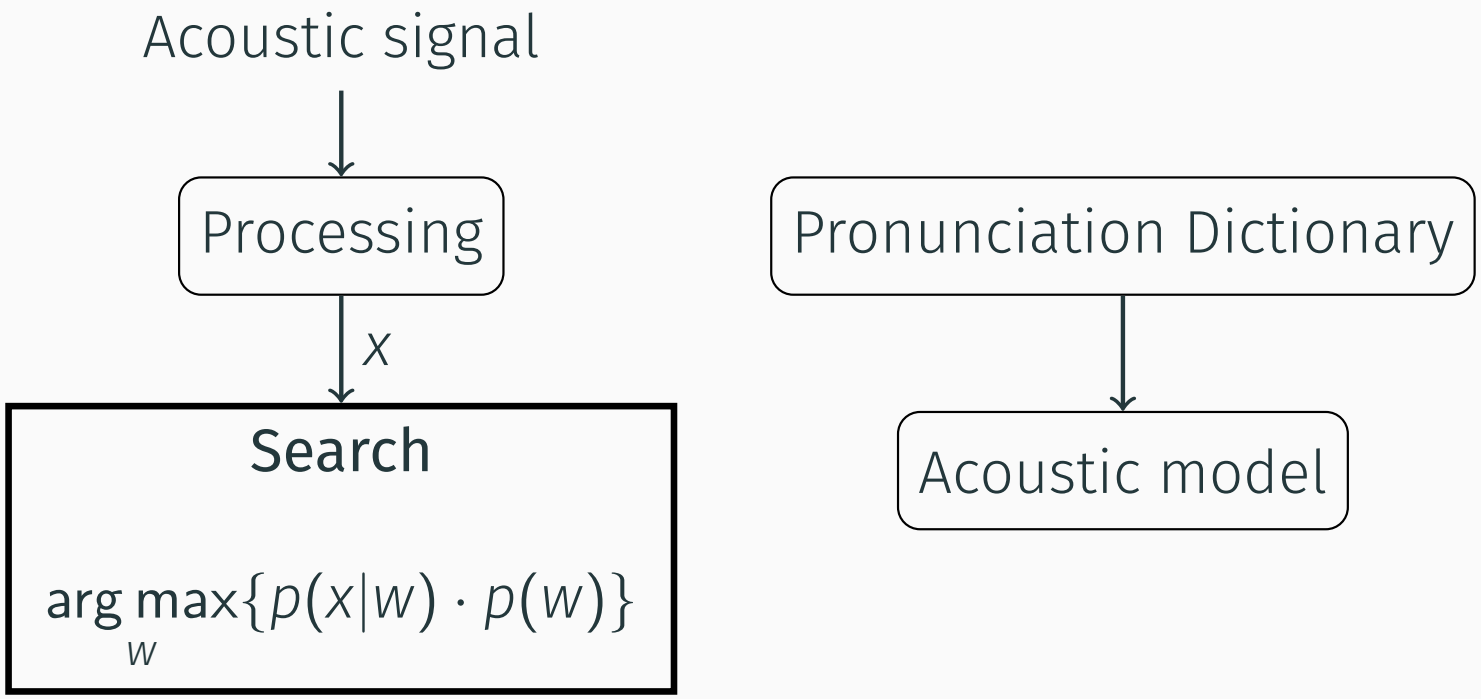
Statistical SR[1]



Statistical SR[1]



Statistical SR[1]

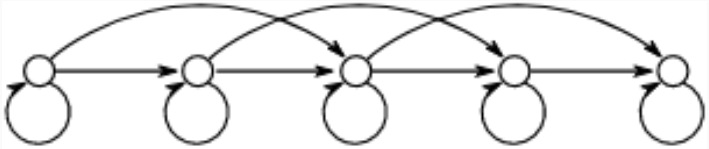


Statistical SR[1] – Acoustic Model

Goal: $p(x|w)$ – Given a word sequence w, find the probability of observing the feature vector x.

- Use the pronunciation dictionary to create ”subwords” and map them to a Hidden Markov Model

$$p(\text{subwords} \mid \text{unknown Markov process})$$

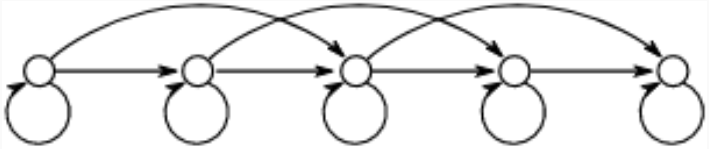


Statistical SR[1] – Acoustic Model

Goal: $p(x|w)$ – Given a word sequence w, find the probability of observing the feature vector x.

- Use the pronunciation dictionary to create ”subwords” and map them to a Hidden Markov Model

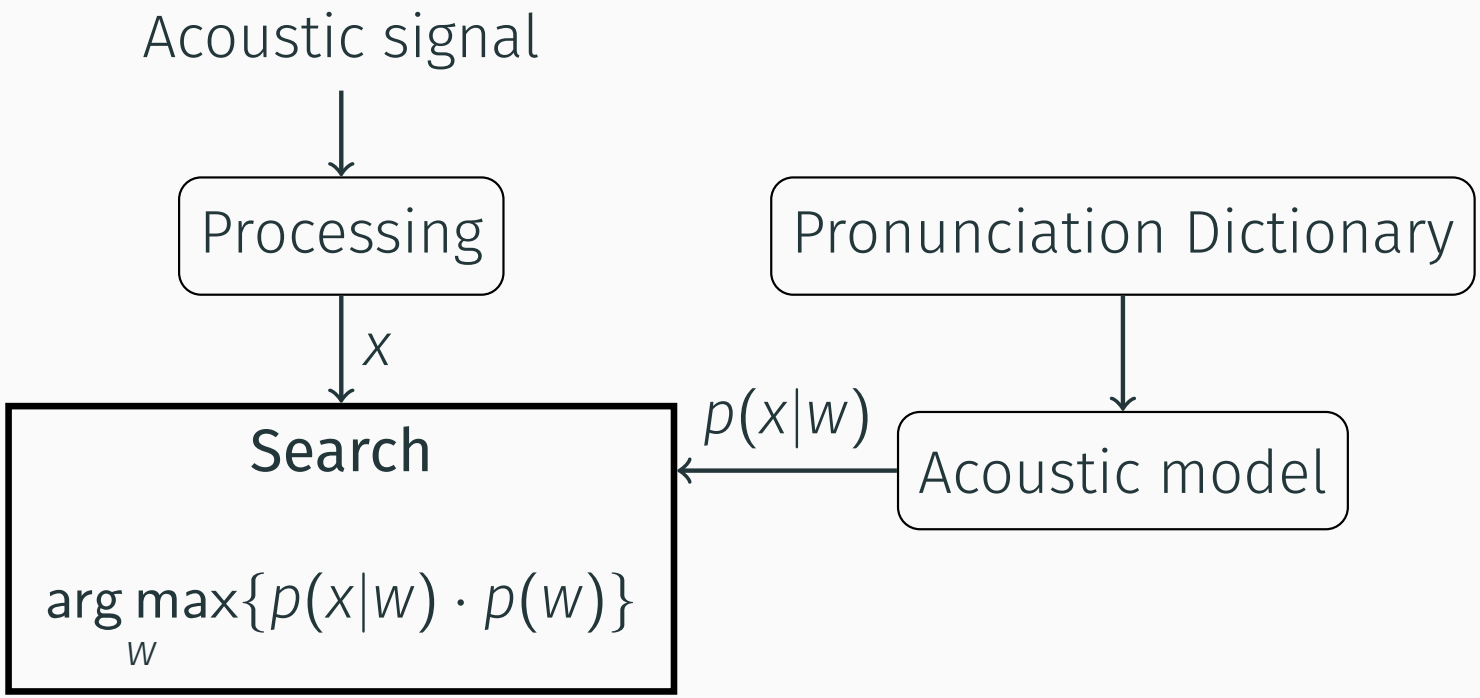
$$p(\text{subwords} \mid \text{unknown Markov process})$$



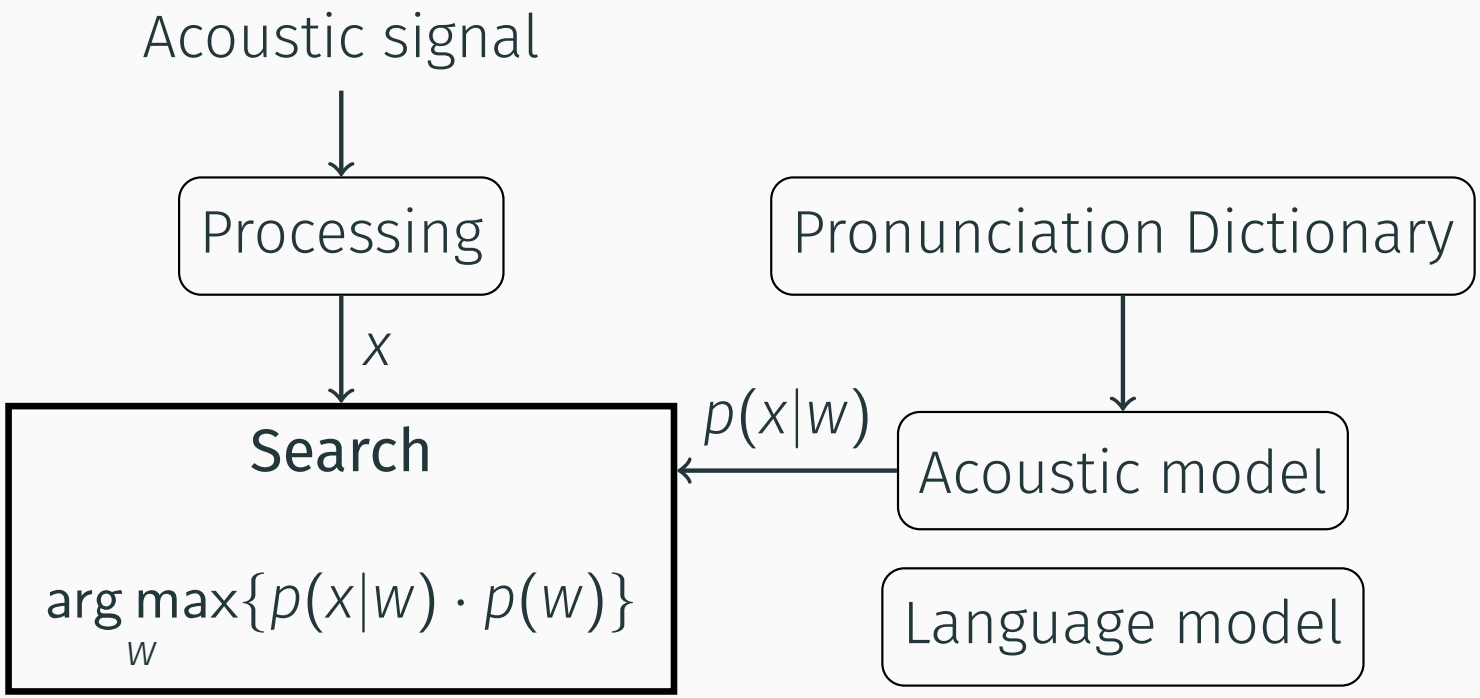
- Markov Property – to calculate probability we only need the **current** state s_t and **previous** state s_{t-1}

$$p(x|w) \approx \max_{s} \prod_{i=1}^{k} p(x_i|s_i, w) \cdot p(s_i|s_{i-1}, w)$$

Statistical SR[1]

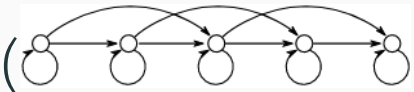


Statistical SR[1]



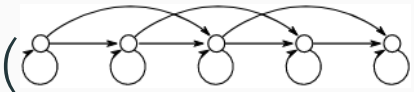
Statistical SR[1] – Language Model

Goal: $p(w)$ – Find the probability of occurrence of a given word sequence w.

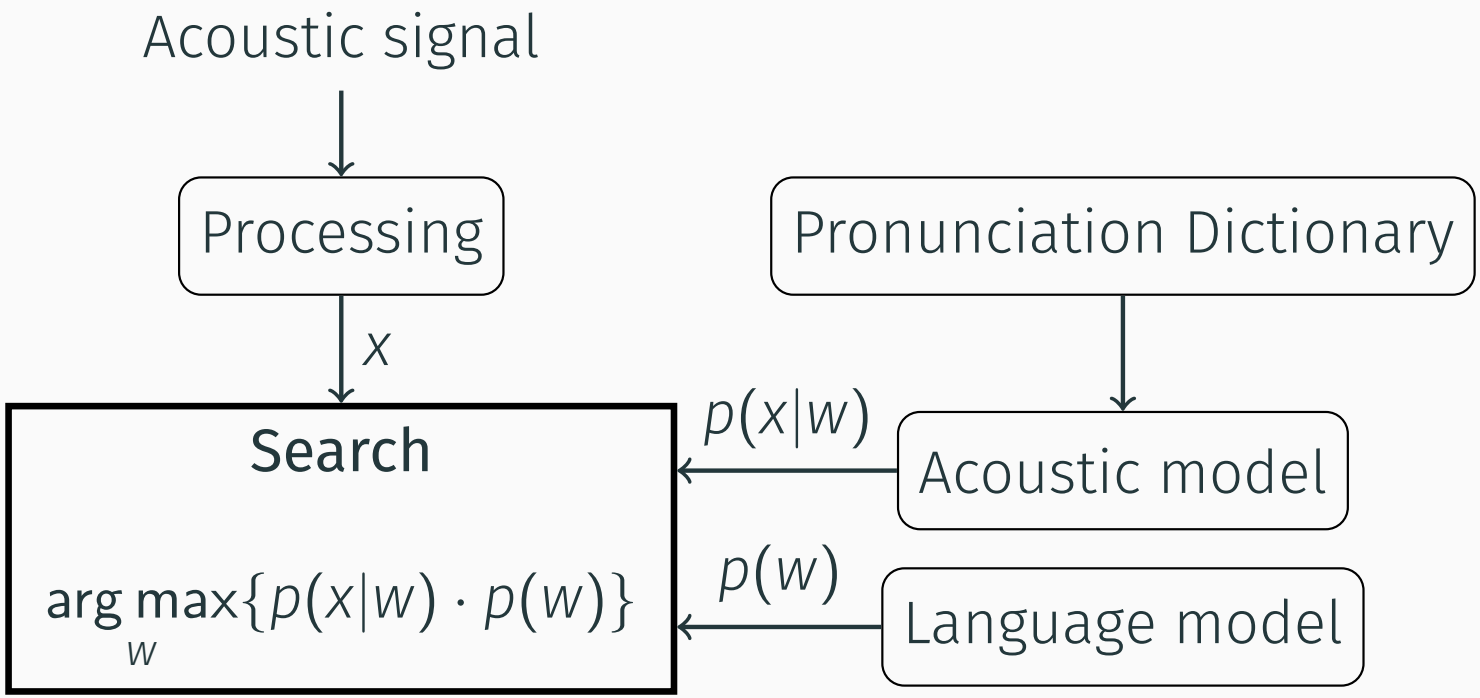
- Represent ”words” using Markov processes
  ( states and probabilities given!)

Statistical SR[1] – Language Model

Goal: $p(w)$ – Find the probability of occurrence of a given word sequence w.

- Represent ”words” using Markov processes ( states and probabilities given!)
- The word probabilities are calculated by using negative logarithm ”scores”

Statistical SR[1]



Statistical SR[1] – Search (Decoding)

Goal: w_{opt} – Find the optimal set of words given the conditional feature probability (AM) and word probability (LM).

- Combine AM and LM information into a single *search space* graph where probabilities are represented as negative logarithm scores

$$w_{opt} = \arg\max\{p(x|w) \cdot p(w)\} = \arg\min_{w}\{LM_{score} + AM_{score}\}$$

Statistical SR[1] – Search (Decoding)

Goal: w_{opt} – Find the optimal set of words given the conditional feature probability (AM) and word probability (LM).

- Combine AM and LM information into a single *search space* graph where probabilities are represented as negative logarithm scores

$$w_{opt} = \arg\max\{p(x|w) \cdot p(w)\} = \arg\min_{w}\{LM_{score} + AM_{score}\}$$

- Perform beam search, a greedy graph search algorithm

Statistical SR[1] – Search (Decoding)

Goal: w_{opt} – Find the optimal set of words given the conditional feature probability (AM) and word probability (LM).

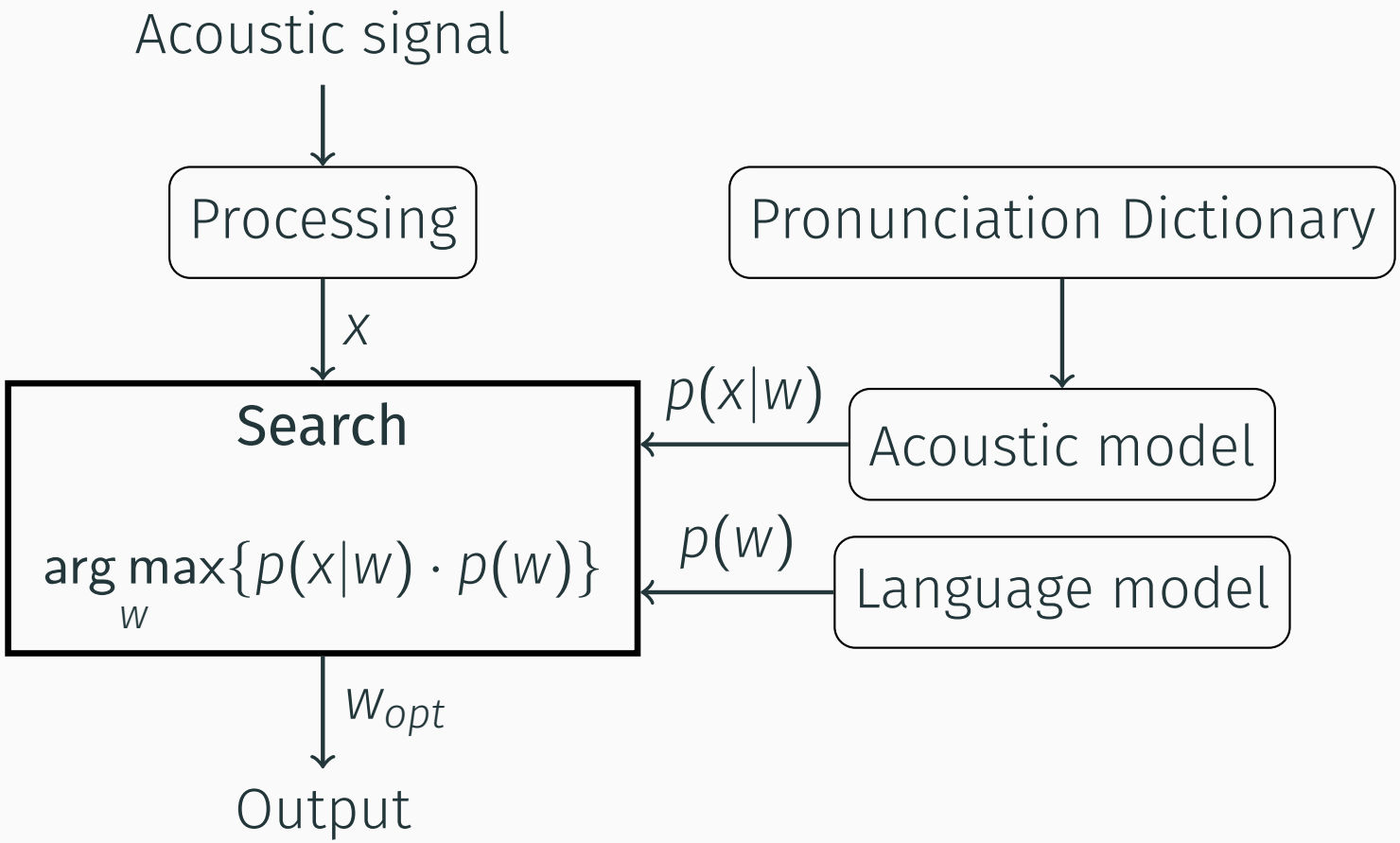
- Combine AM and LM information into a single *search space* graph where probabilities are represented as negative logarithm scores

$$w_{opt} = \arg\max\{p(x|w) \cdot p(w)\} = \arg\min_{w}\{LM_{score} + AM_{score}\}$$

- Perform beam search, a greedy graph search algorithm
  - Scores need to be normalized[3] otherwise shorter sentences will have much higher scores compared to longer sentences at sentence level:

$$\text{length penalty} = \frac{(5 + |S|)^{\alpha}}{(5 + 1)^{\alpha}}$$

Statistical SR[1]

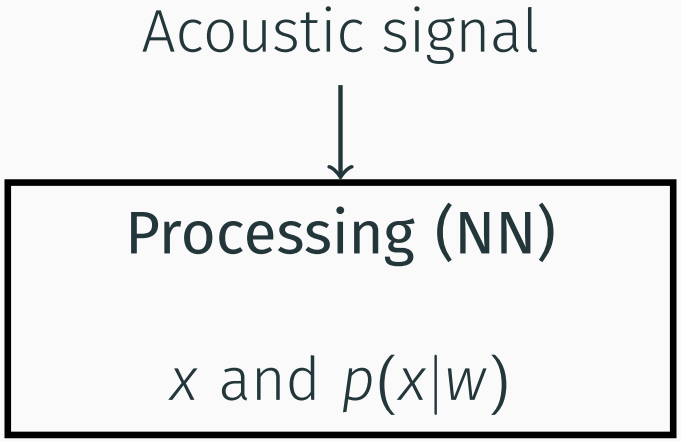


End-to-End SR^2

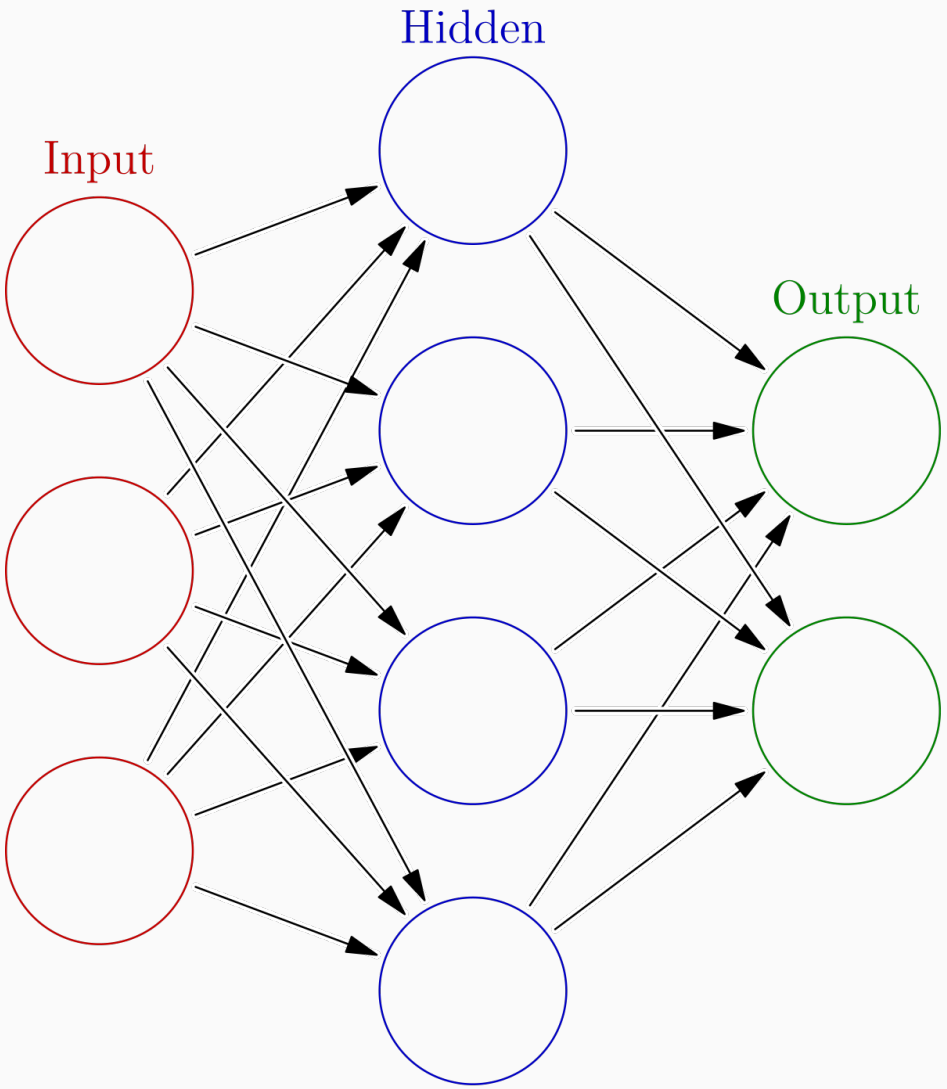
End-to-End SR^2

Acoustic signal

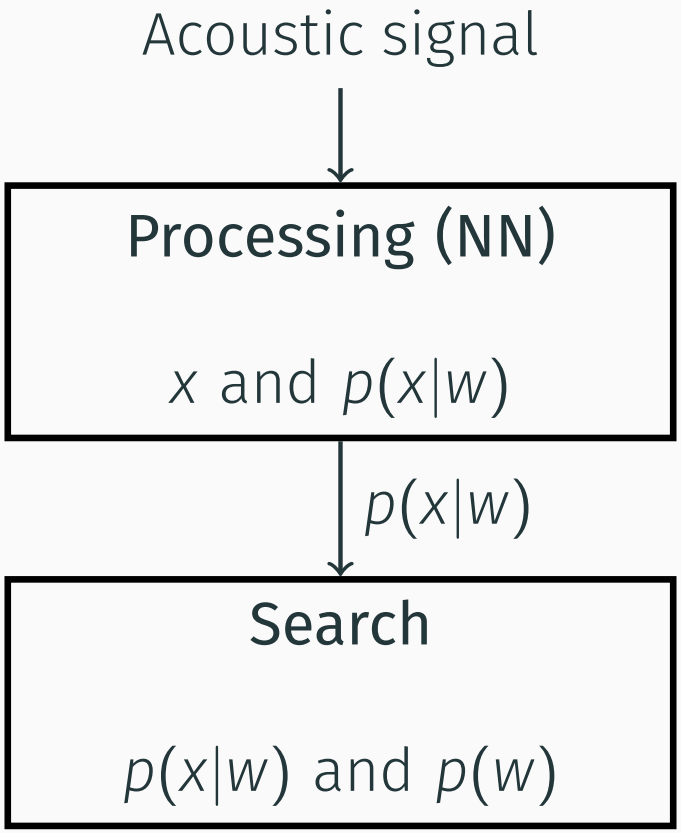
End-to-End SR^2



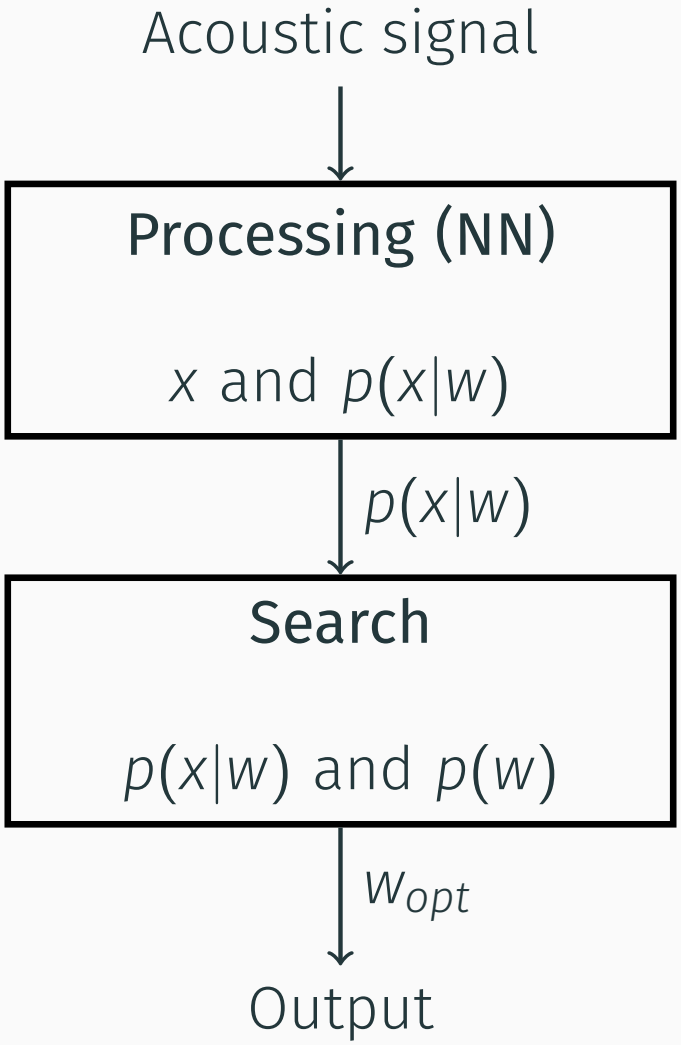
End-to-End SR^2



End-to-End SR[2]



End-to-End SR²



Statistical vs. End-to-End SR

- Despite the immense complexity, statistical SR is interpretable (confidence intervals for words)

Statistical vs. End-to-End SR

- Despite the immense complexity, statistical SR is interpretable (confidence intervals for words)
- However, end-to-end SR is much easier to implement (not necessarily train) and rivals the speed of statistical SR

Statistical vs. End-to-End SR

- Despite the immense complexity, statistical SR is interpretable (confidence intervals for words)
- However, end-to-end SR is much easier to implement (not necessarily train) and rivals the speed of statistical SR
- Currently, end-to-end SR seems to be replacing the traditional statistical approach.

References

References

D. Nolden, H. Ney, and J.-L. Gauvain.
Progress in decoding for large vocabulary continuous speech recognition.
PhD thesis, 2017.

D. Palaz.
Towards end-to-end speech recognition.
PhD thesis, 2016.

Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey, et al.
Google's neural machine translation system: Bridging the gap between human and machine translation.
arXiv preprint arXiv:1609.08144, 2016.

Thank you!