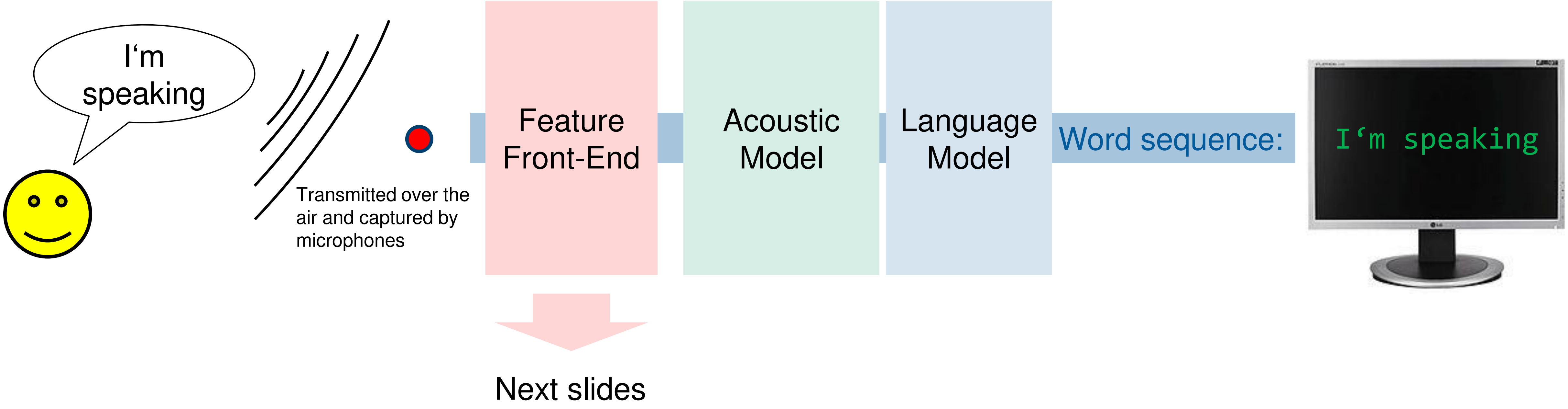
Automatic Speech Recognition

Introduction

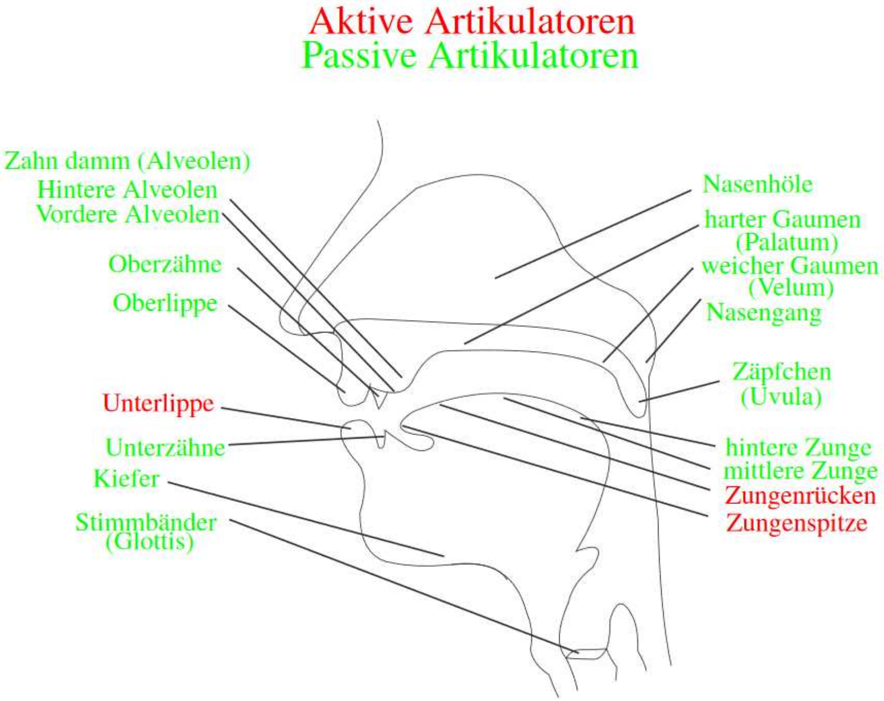
Automatic Speech Recognition, is the technology that allows human beings to use their voices to speak with a computer interface in a way that, in its most sophisticated variations, resembles normal human conversation.



Subfield of articulatory phonetics

Vowel tract & classification of sounds

The vocal tract can be well described as an all-pole filter, which can be useful, for example, for the analysis or synthesis of speech signals. The speech organs that play a special role in sound production or shaping are called articulators. A distinction is made between the more or less consciously influenced articulators and those that are only used, or between active and passive articulators. In order to describe the many, different sounds of the human language, one needs first a smallest unit, which can serve as basis for a description alphabet. In phonetics, this smallest unit is called a sound or a phon.

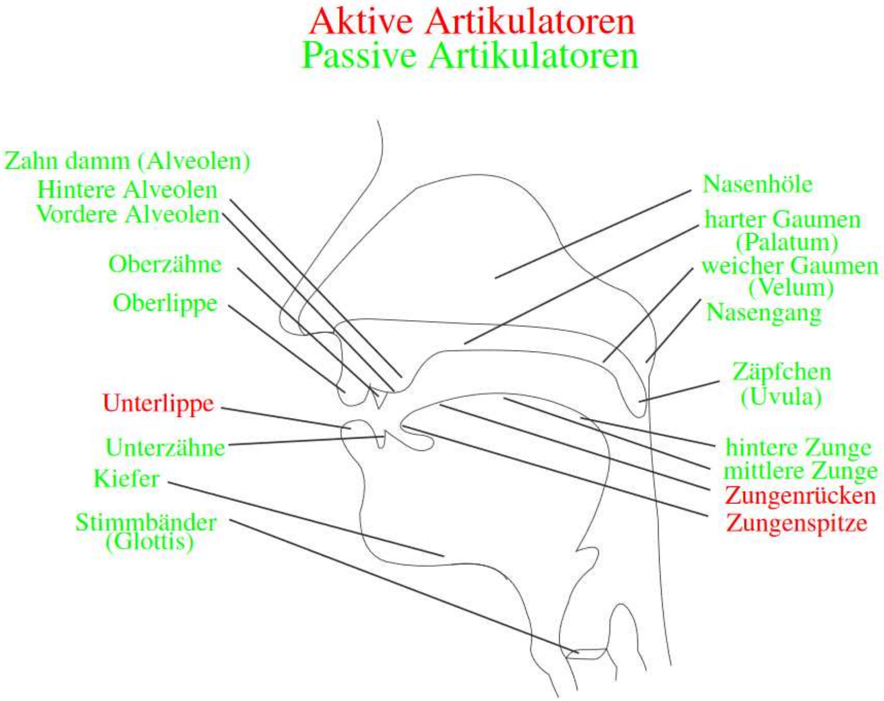


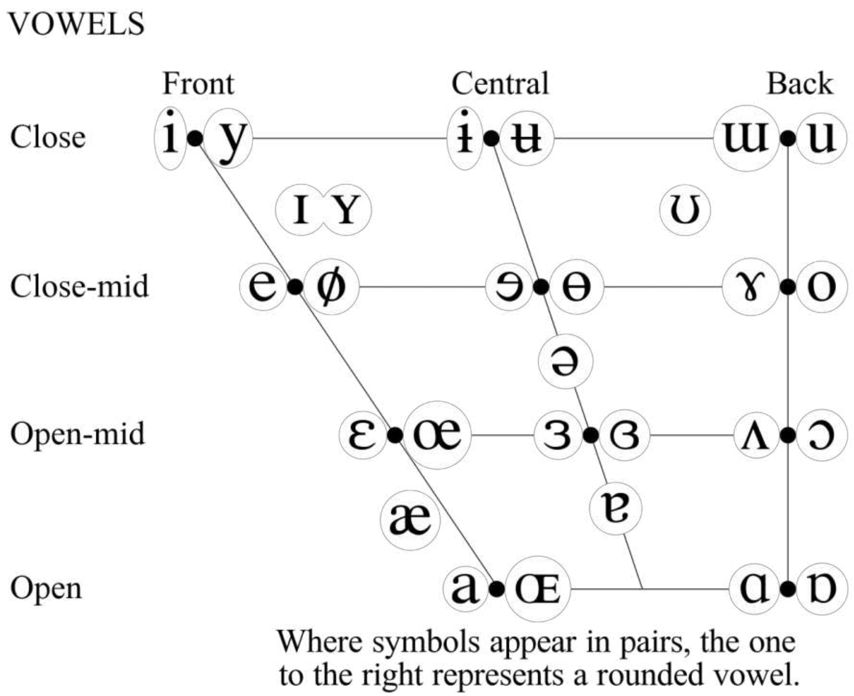
 https://www.ei.ruhr-uni-bochum.de/media/ei/lehrmaterialien/spracherkennung/d0909549b9003dc1defe7f7a960ce6624d1de7f9/SkriptASE2017b.pdf

Subfield of articulatory phonetics

Vowel tract & classification of sounds

The vocal tract can be well described as an all-pole filter, which can be useful, for example, for the analysis or synthesis of speech signals. The speech organs that play a special role in sound production or shaping are called articulators. A distinction is made between the more or less consciously influenced articulators and those that are only used, or between active and passive articulators. In order to describe the many, different sounds of the human language, one needs first a smallest unit, which can serve as basis for a description alphabet. In phonetics, this smallest unit is called a sound or a phon.



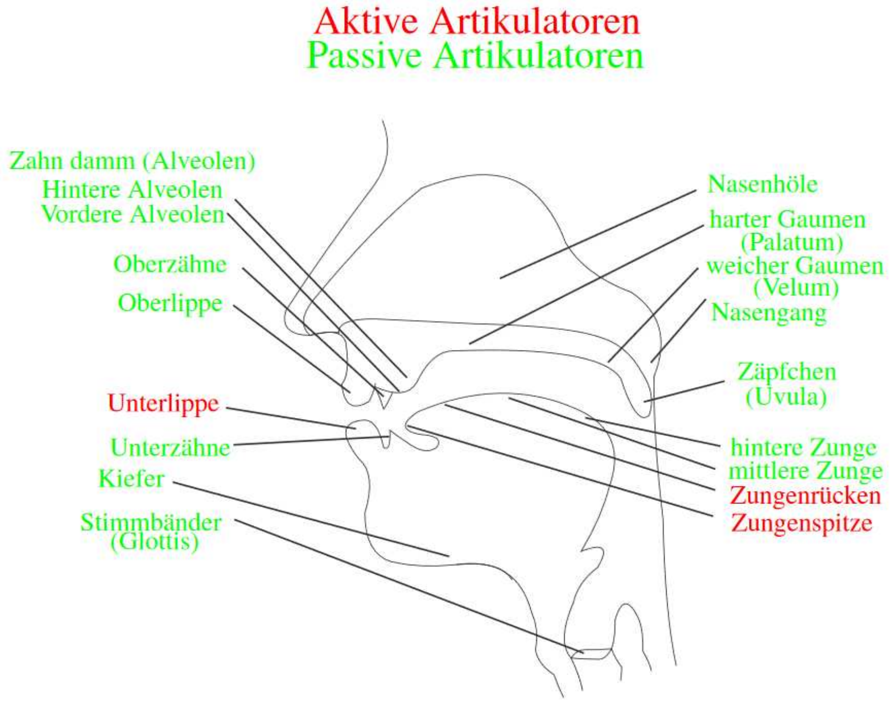


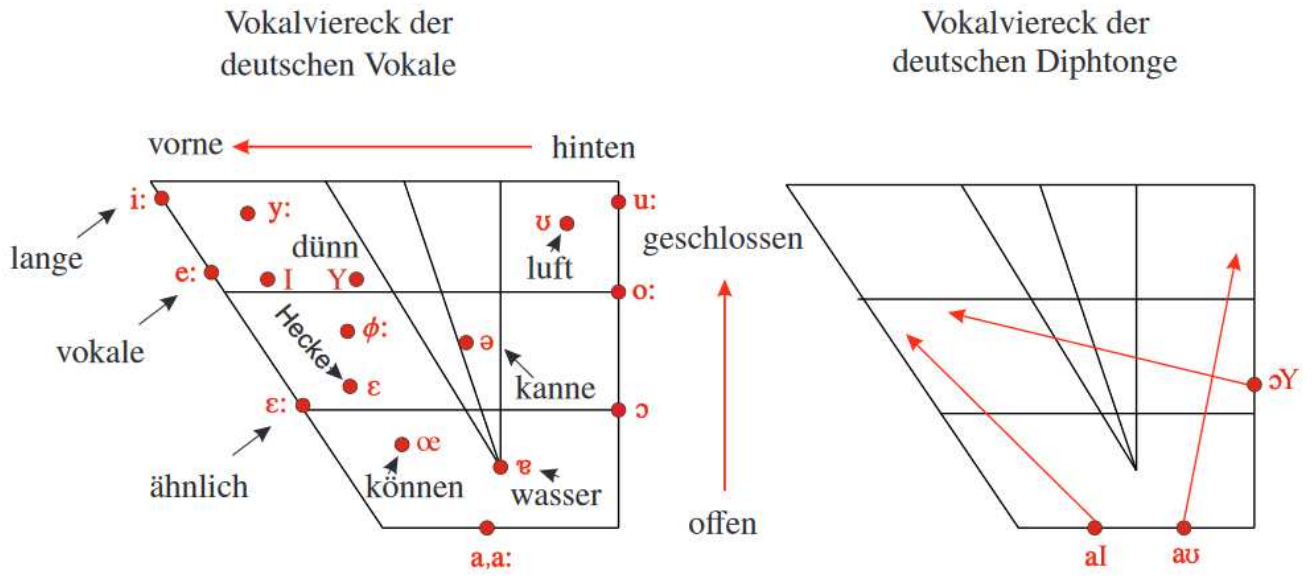
 https://www.ei.ruhr-uni-bochum.de/media/ei/lehrmaterialien/spracherkennung/d0909549b9003dc1defe7f7a960ce6624d1de7f9/SkriptASE2017b.pdf

Subfield of articulatory phonetics

Vowel tract & classification of sounds

The vocal tract can be well described as an all-pole filter, which can be useful, for example, for the analysis or synthesis of speech signals. The speech organs that play a special role in sound production or shaping are called articulators. A distinction is made between the more or less consciously influenced articulators and those that are only used, or between active and passive articulators. In order to describe the many, different sounds of the human language, one needs first a smallest unit, which can serve as basis for a description alphabet. In phonetics, this smallest unit is called a sound or a phon.



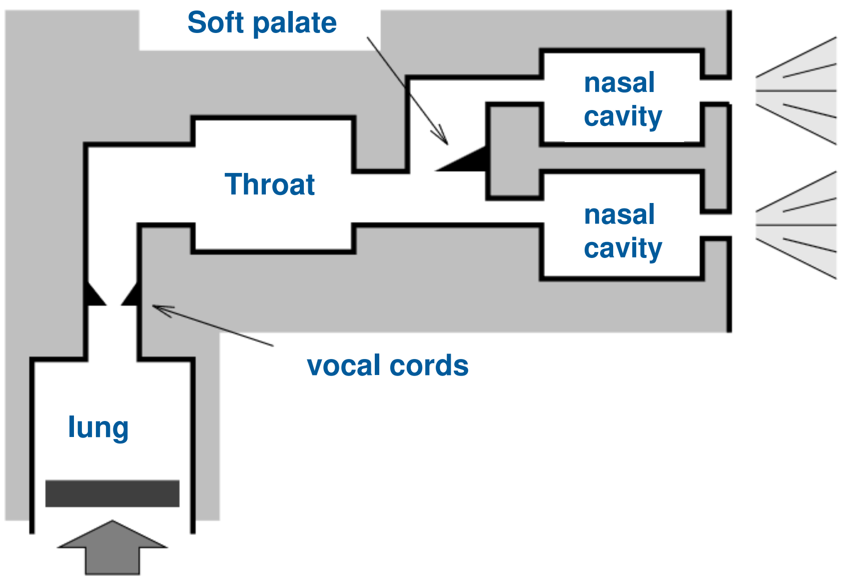


 https://www.ei.ruhr-uni-bochum.de/media/ei/lehrmaterialien/spracherkennung/d0909549b9003dc1defe7f7a960ce6624d1de7f9/SkriptASE2017b.pdf

Subfield of articulatory phonetics

Physiologically motivated model of speech generation

To describe speech generation mathematically, the model in the lower left image is often used. Here, the lung serves as the source that provides the airflow for all further processes. The vocal cords determine whether the sound is to be voiced or unvoiced. In the case of unvoiced sounds, the vocal cords are so far apart that they are not influenced too much by the passing air stream; in the case of voiced sounds, they lie against each other and are moved apart at regular intervals by the air stream, thus causing them to vibrate. The frequency of this oscillation is also referred to as the fundamental frequency.

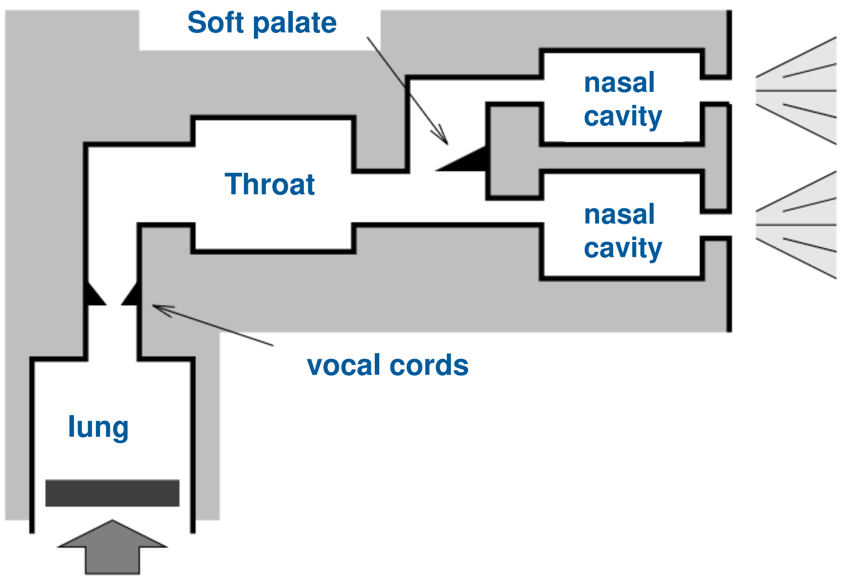


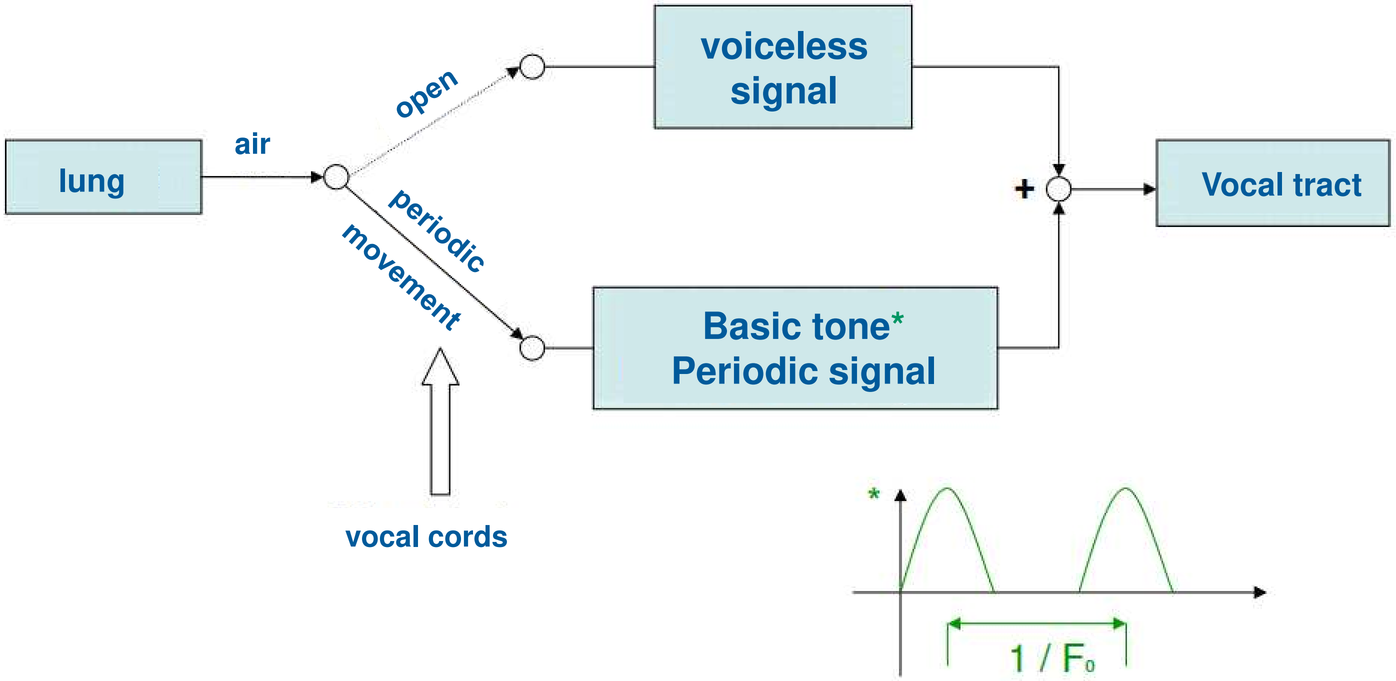
 https://www.ei.ruhr-uni-bochum.de/media/ei/lehrmaterialien/spracherkennung/d0909549b9003dc1defe7f7a960ce6624d1de7f9/SkriptASE2017b.pdf

Subfield of articulatory phonetics

Physiologically motivated model of speech generation

To describe speech generation mathematically, the model in the lower left image is often used. Here, the lung serves as the source that provides the airflow for all further processes. The vocal cords determine whether the sound is to be voiced or unvoiced. In the case of unvoiced sounds, the vocal cords are so far apart that they are not influenced too much by the passing air stream; in the case of voiced sounds, they lie against each other and are moved apart at regular intervals by the air stream, thus causing them to vibrate. The frequency of this oscillation is also referred to as the fundamental frequency.



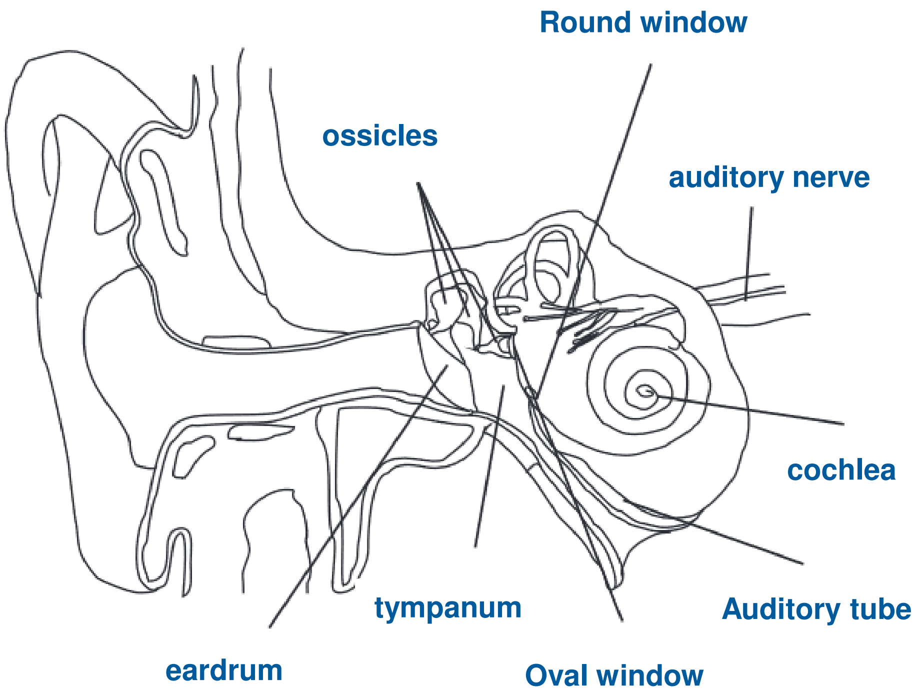


 https://www.ei.ruhr-uni-bochum.de/media/ei/lehrmaterialien/spracherkennung/d0909549b9003dc1defe7f7a960ce6624d1de7f9/SkriptASE2017b.pdf

Auditory apparatus

Structure of the auditory system

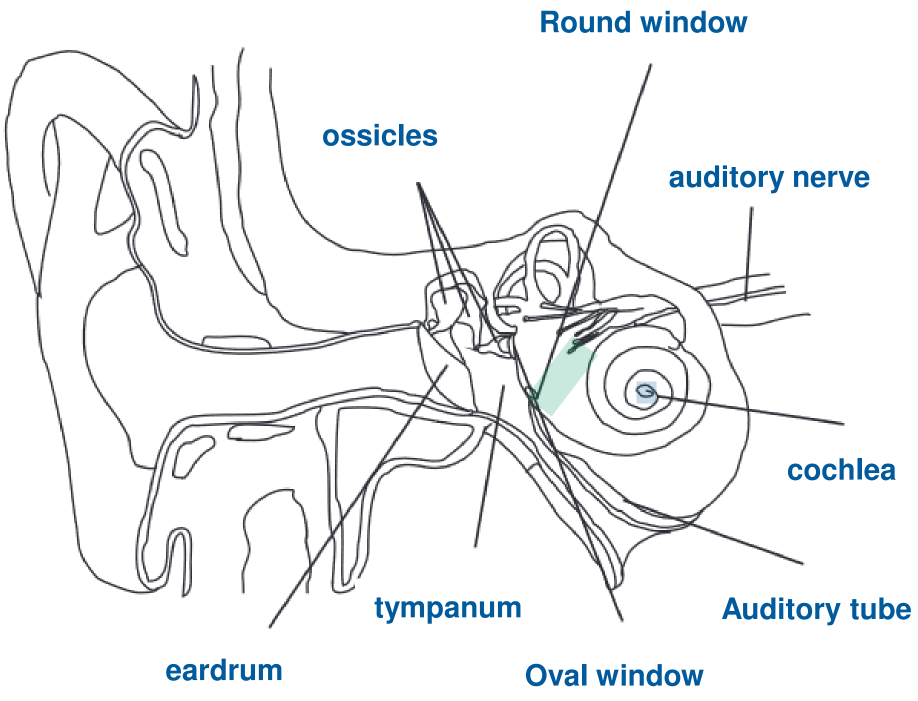
The outer ear, for its part, consists of the auricle, whose directional characteristics, among other things, make it easier to focus on sounds from a particular direction of incidence, the auditory canal, which primarily keeps foreign bodies out, and is bounded by the eardrum, which is stimulated to vibrate by sound waves. In the middle ear, the three ossicles, malleus, incus, and stapes, effect impedance matching, which is necessary because the sound resistance of the fluid-filled inner ear is much greater than that of air, so that without appropriate mechanical transduction, sound would have no appreciable effect on the inner ear

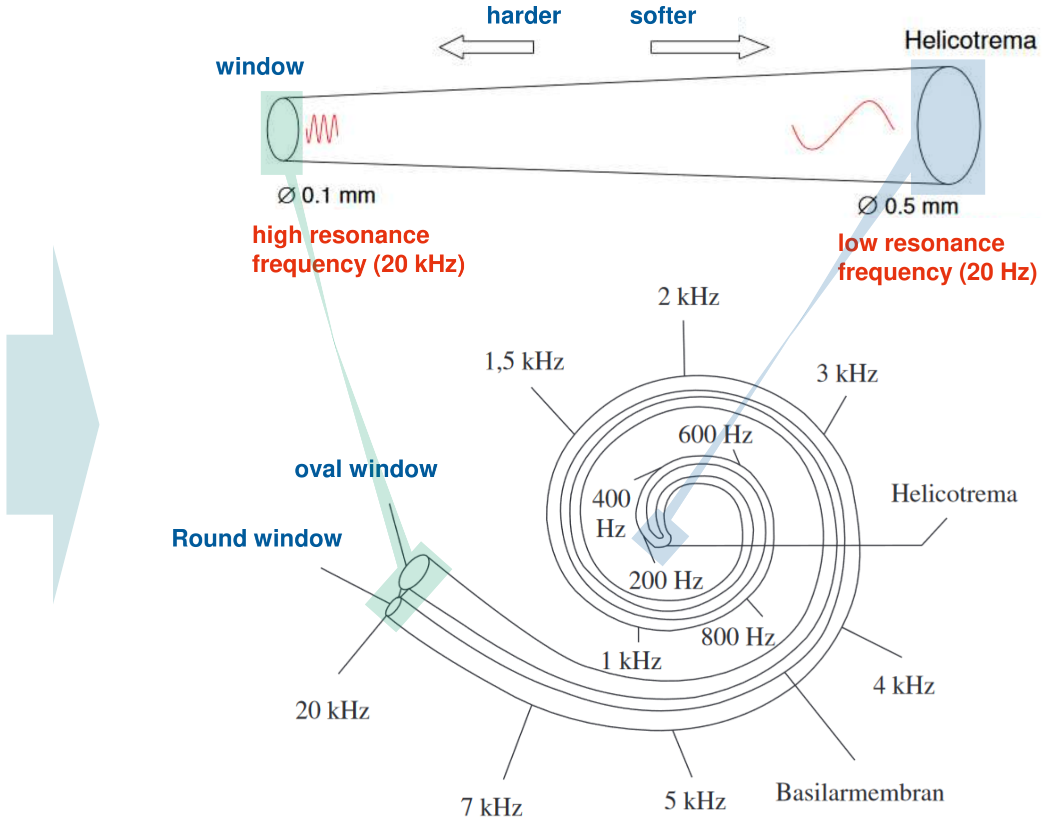


Auditory apparatus

Structure of the auditory system

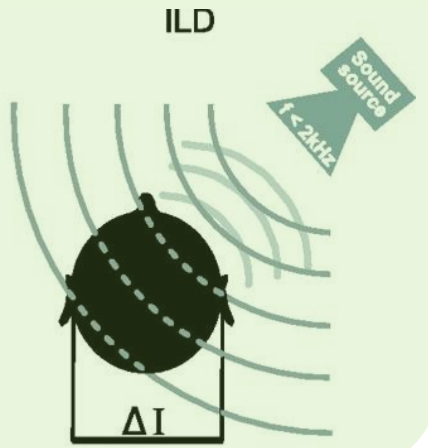
The outer ear, for its part, consists of the auricle, whose directional characteristics, among other things, make it easier to focus on sounds from a particular direction of incidence, the auditory canal, which primarily keeps foreign bodies out, and is bounded by the eardrum, which is stimulated to vibrate by sound waves. In the middle ear, the three ossicles, malleus, incus, and stapes, effect impedance matching, which is necessary because the sound resistance of the fluid-filled inner ear is much greater than that of air, so that without appropriate mechanical transduction, sound would have no appreciable effect on the inner ear

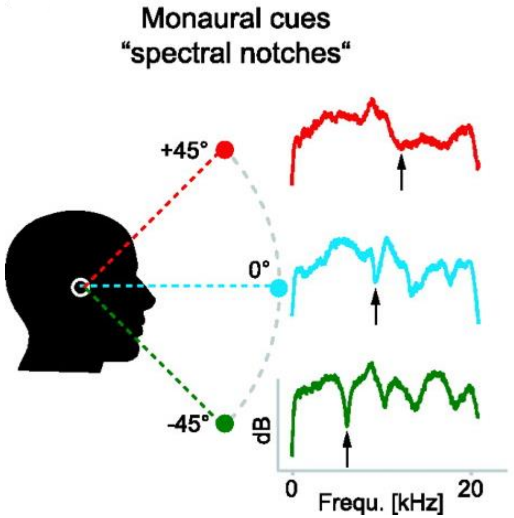


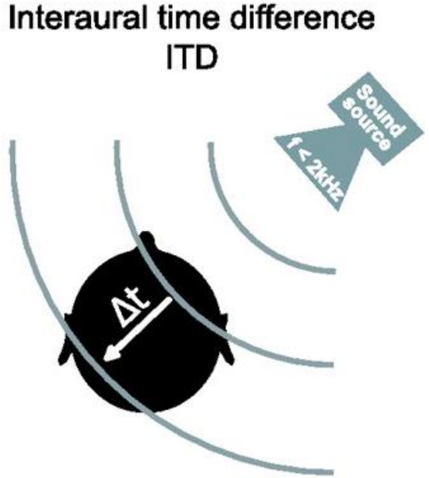


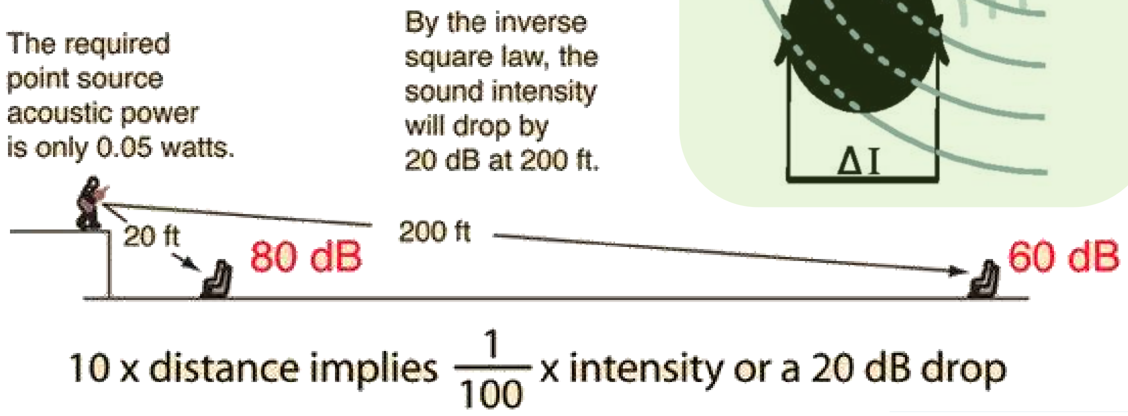
Feature Extraction for Speech Applications

N-Channel Audio Capturing









Feature Extraction for Speech Applications

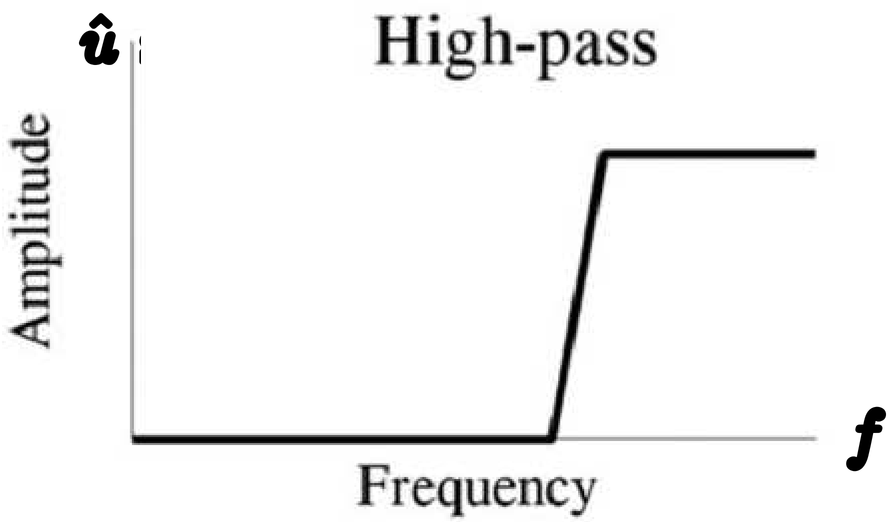
N-Channel Audio Capturing

High pass filter

$$u(t) = \hat{u}\,\sin(\omega t) = \hat{u}\,\sin(2\pi f t) = \hat{u}\,\sin\frac{2\pi t}{T}$$ with $f = \frac{1}{T}$

Example:

$f = 50\ \mathrm{Hz}$



Remove low frequencies from the signal

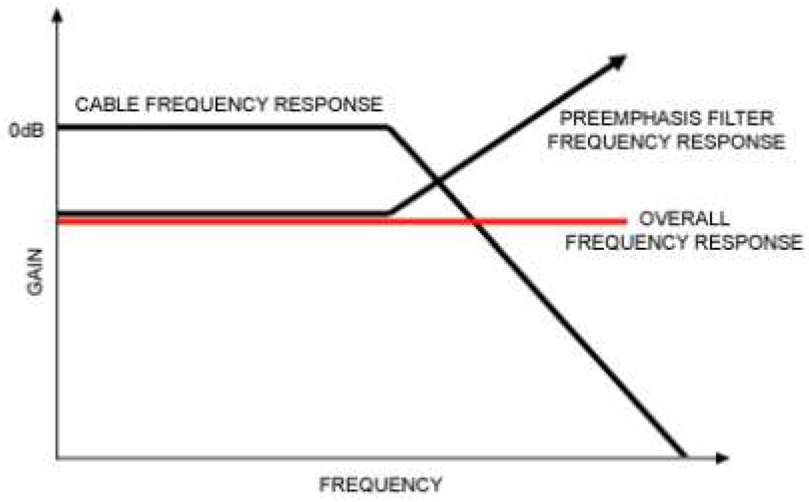
https://en.wikipedia.org/wiki/High-pass_filter

Feature Extraction for Speech Applications

N-Channel Audio Capturing

High pass filter

Pre-emphasis



$$\tau = \frac{1}{2\pi \cdot f_c}$$

https://www.eeweb.com/introduction-to-preemphasis-and-equalization-in-maxim-gmsl-serdes-devices/

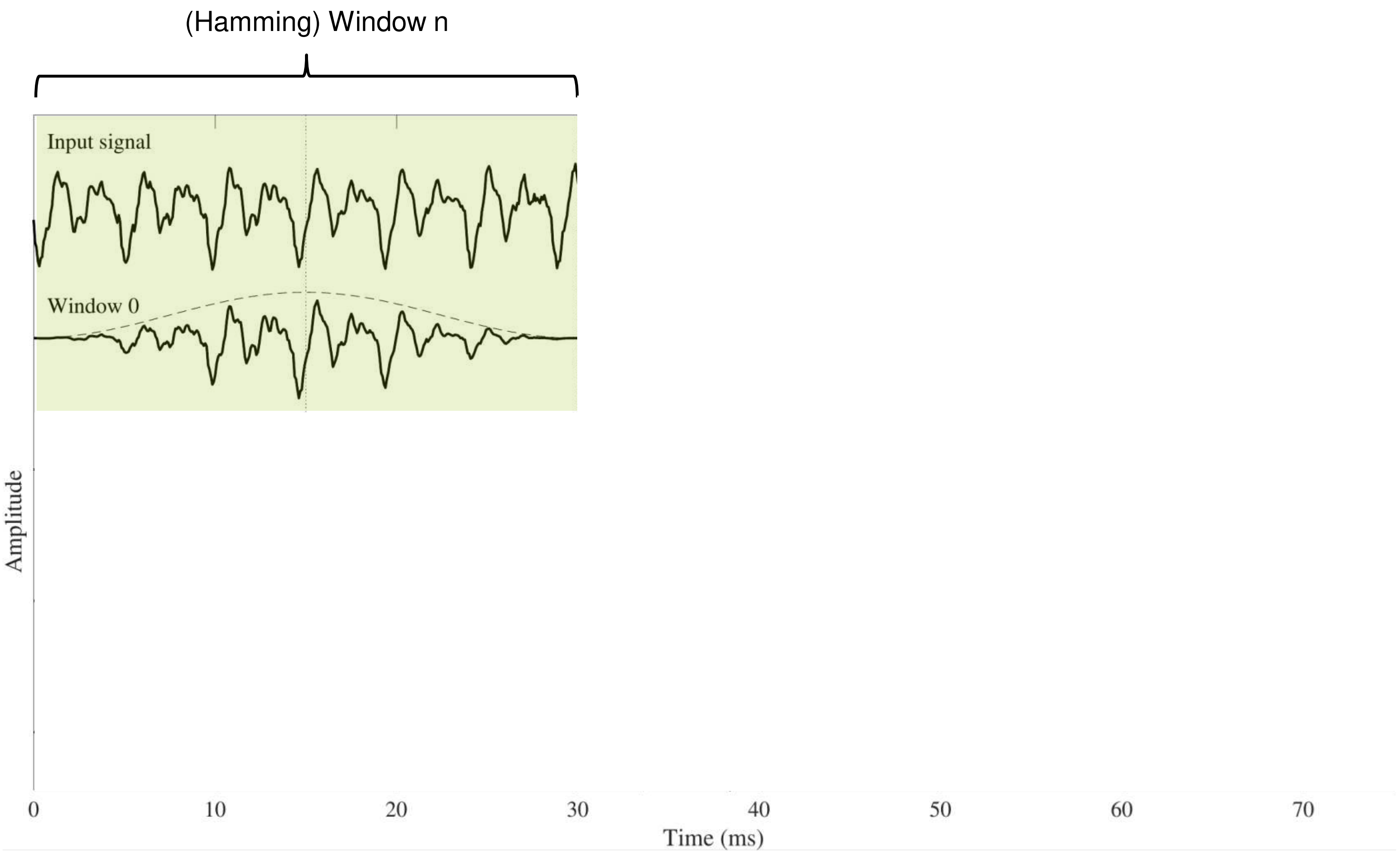
Feature Extraction for Speech Applications

N-Channel Audio Capturing

High pass filter

Pre-emphasis

Windowing



Compute Feature Vector

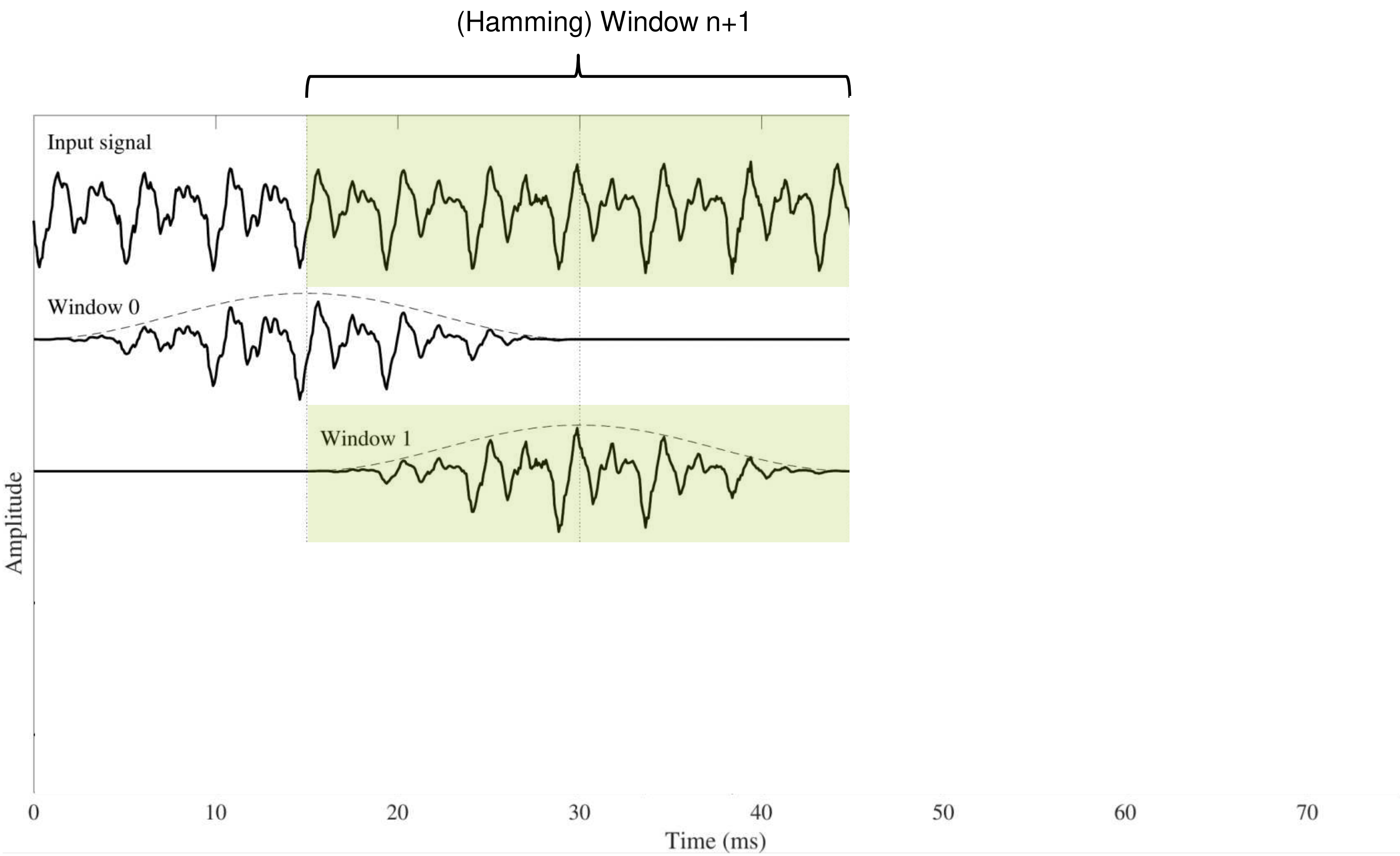
Feature Extraction for Speech Applications

N-Channel Audio Capturing

High pass filter

Pre-emphasis

Windowing



Compute Feature Vector

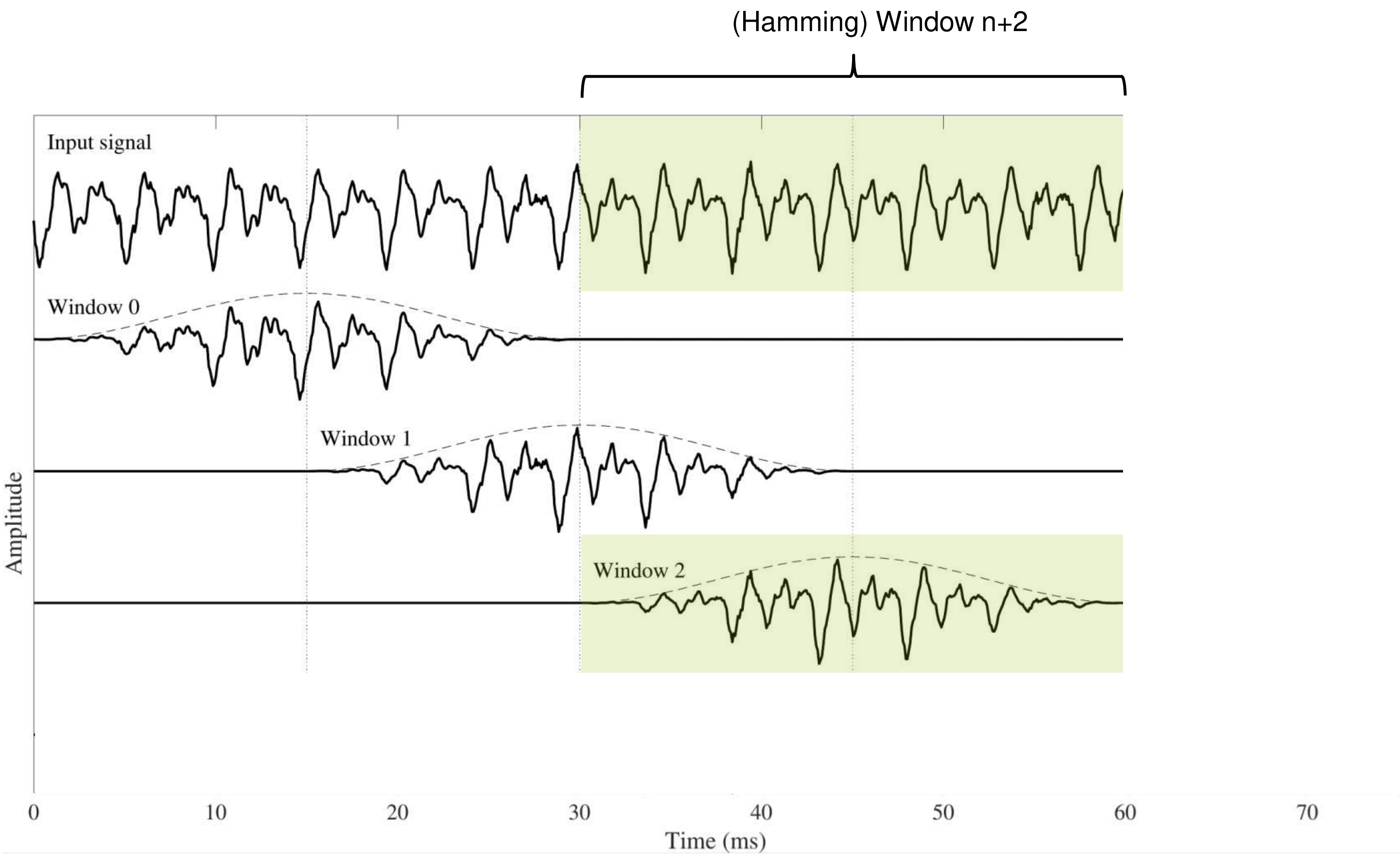
Feature Extraction for Speech Applications

N-Channel Audio Capturing

High pass filter

Pre-emphasis

Windowing



Compute Feature Vector

Feature Extraction for Speech Applications

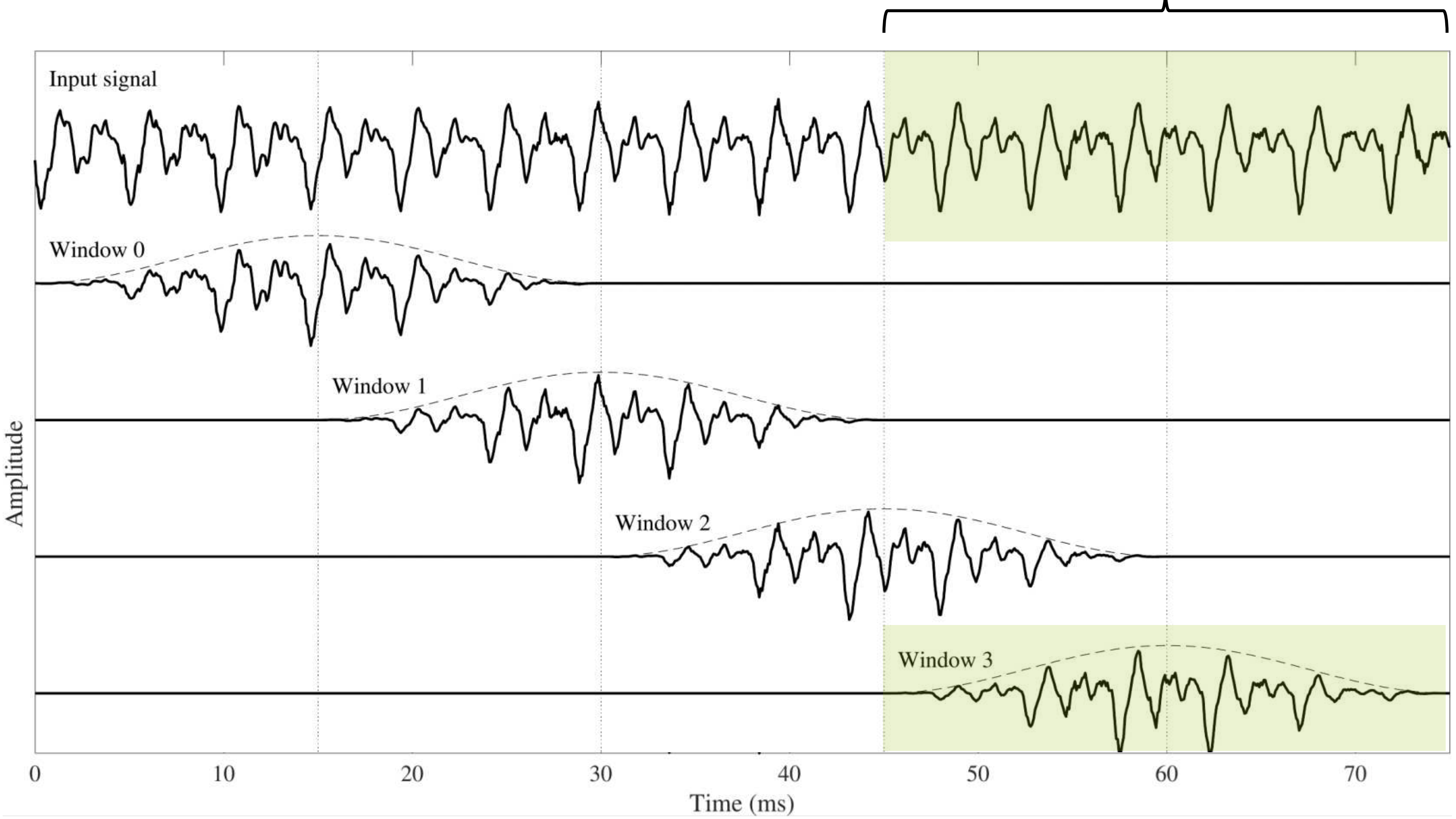
N-Channel Audio Capturing

High pass filter

Pre-emphasis

Windowing

(Hamming) Window n+3



Compute Feature Vector

Feature Extraction for Speech Applications

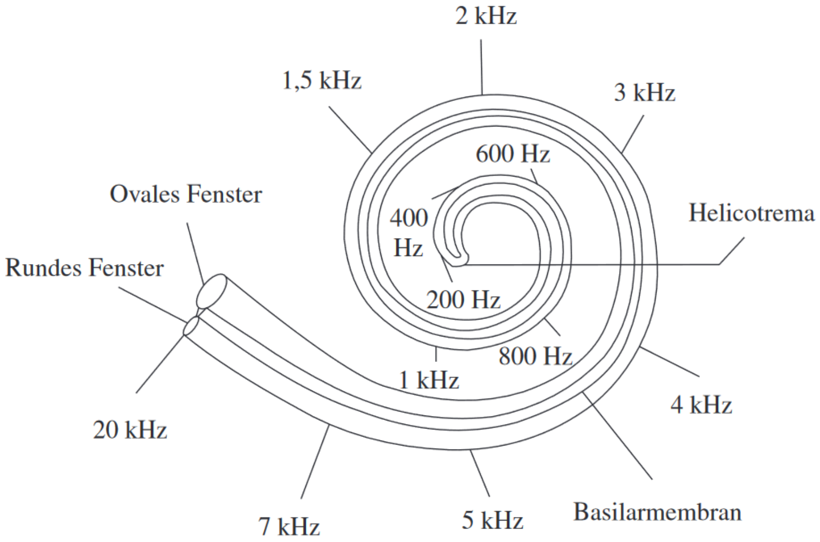
N-Channel Audio Capturing

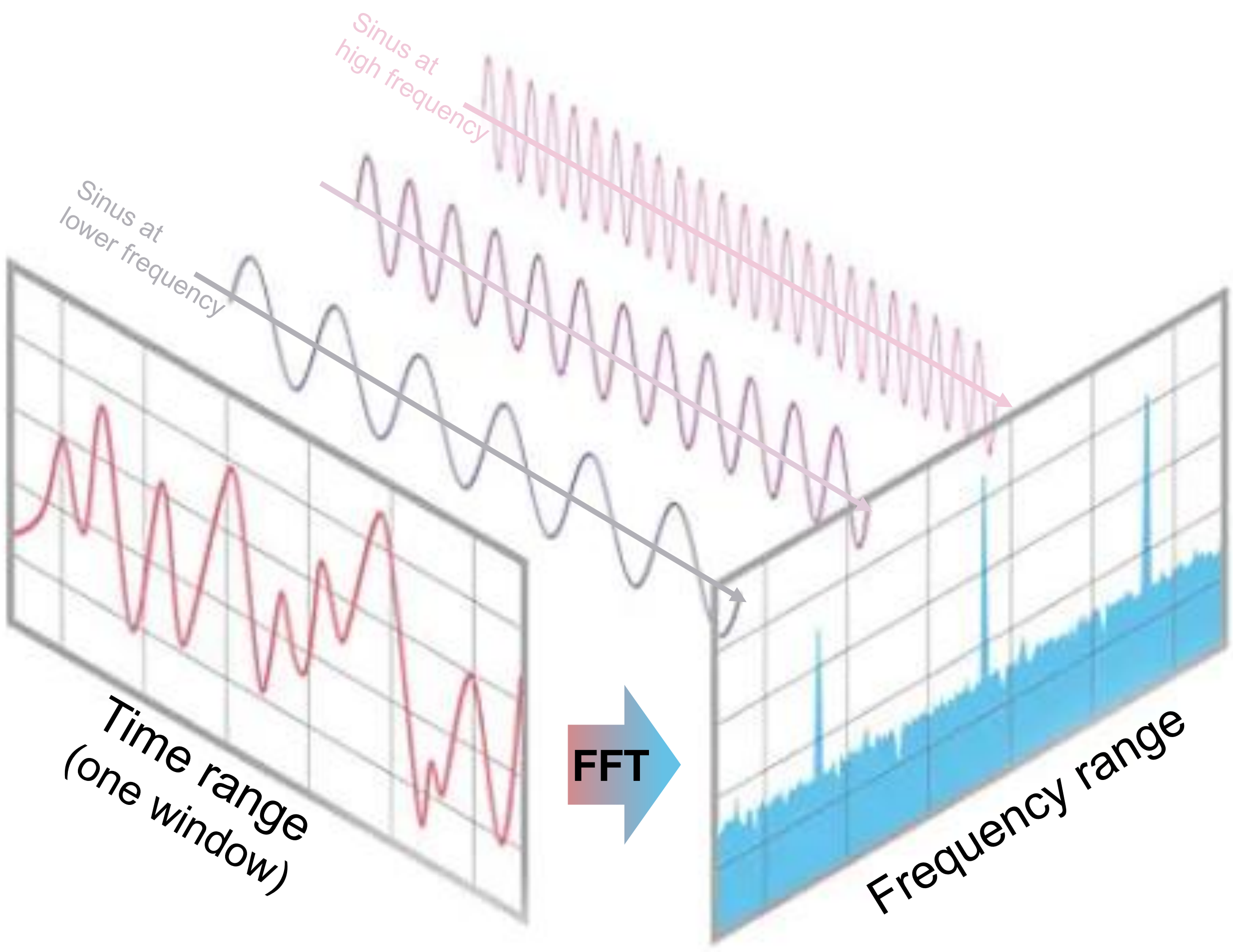
High pass filter

Pre-emphasis

Windowing

Fast Fourier Transformation





https://www.wikiwand.com/en/Fast_Fourier_transform

Feature Extraction for Speech Applications

N-Channel Audio Capturing

High pass filter

Pre-emphasis

Windowing

Fast Fourier Transformation

Absolute Value

$$|x| = \begin{cases} x & \text{if } x \geq 0 \\ -x & \text{if } x < 0 \end{cases}$$

LPCC := Linear Predictive Cepstral Coefficient MFCC := Mel Frequency Filter Bank

Feature Extraction for Speech Applications

N-Channel Audio Capturing

High pass filter

Pre-emphasis

Windowing

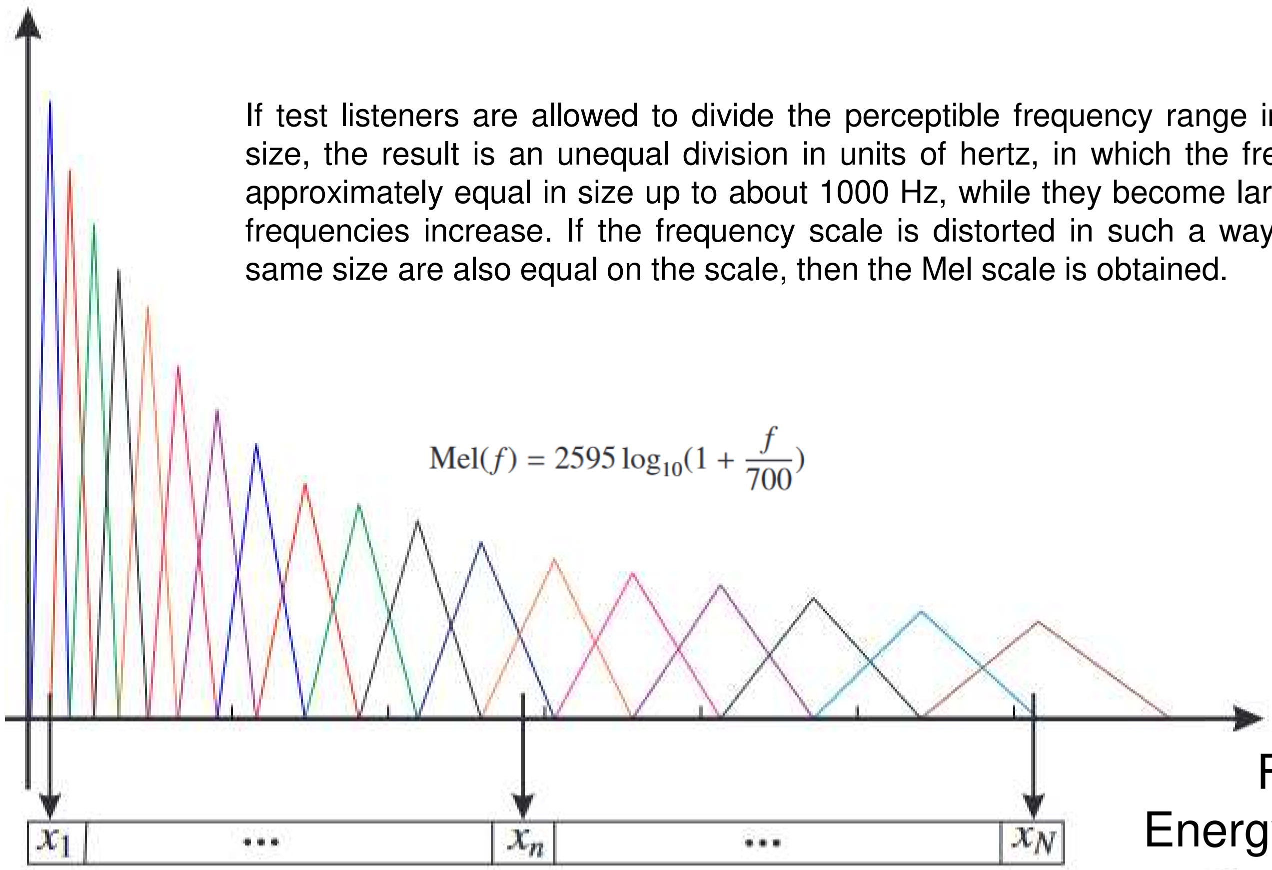
Fast Fourier Transformation

Absolute Value

Mel-Scale Filter bank

If test listeners are allowed to divide the perceptible frequency range into intervals of equal size, the result is an unequal division in units of hertz, in which the frequency intervals are approximately equal in size up to about 1000 Hz, while they become larger and larger as the frequencies increase. If the frequency scale is distorted in such a way that intervals of the same size are also equal on the scale, then the Mel scale is obtained.

$$\mathrm{Mel}(f) = 2595 \log_{10}\left(1 + \frac{f}{700}\right)$$



Frequency

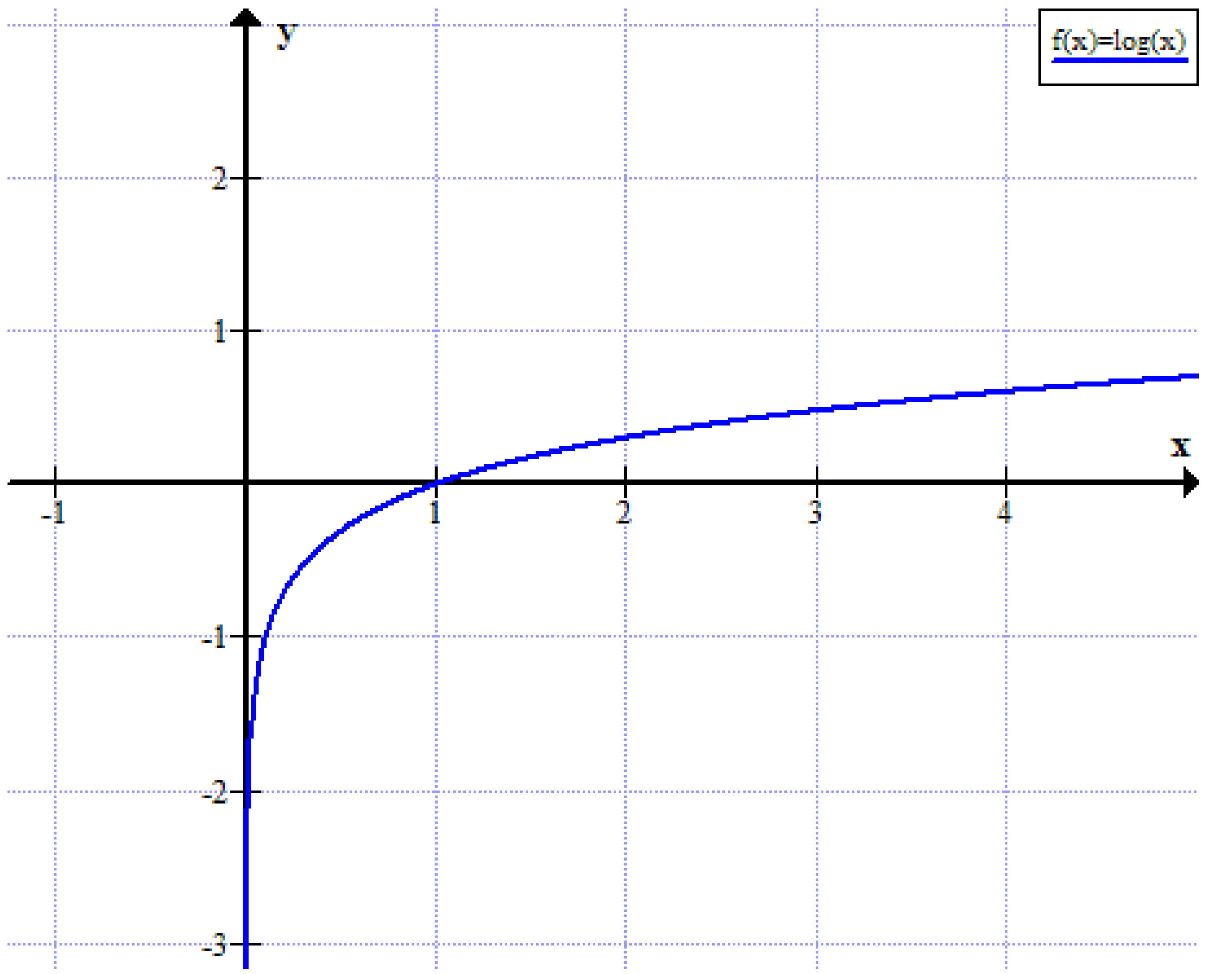
Energy in every band

Stevens, Volkmann, and Newman in 1937

https://en.wikipedia.org/wiki/Mel_scale

Feature Extraction for Speech Applications

- N-Channel Audio Capturing
- High pass filter
- Pre-emphasis
- Windowing
- Fast Fourier Transformation
- Absolute Value
- Mel-Scale Filter bank
- Log-scale

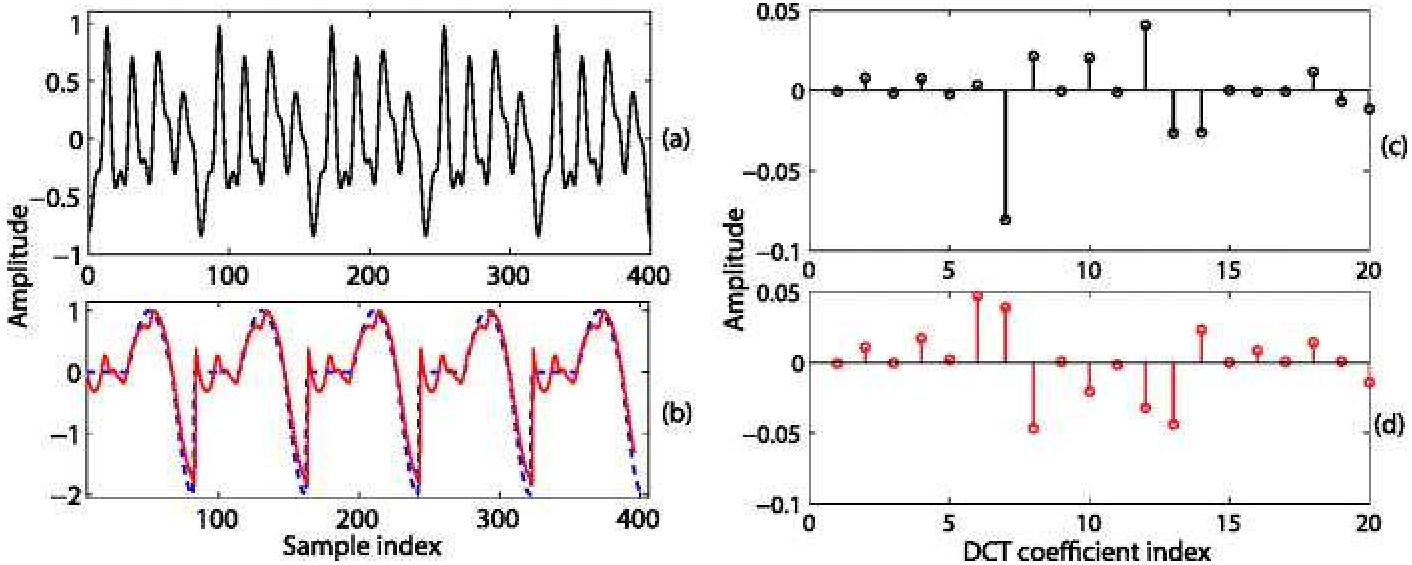


https://en.wikipedia.org/wiki/Logarithmic_scale

Feature Extraction for Speech Applications

- N-Channel Audio Capturing
- High pass filter
- Pre-emphasis
- Windowing
- Fast Fourier Transformation
- Absolute Value
- Mel-Scale Filter bank
- Log-scale
- Discrete Cosine Transformation II

Decorrelation is a general term for any process that is used to reduce autocorrelation within a signal, or cross-correlation within a set of signals, while preserving other aspects of the signal.

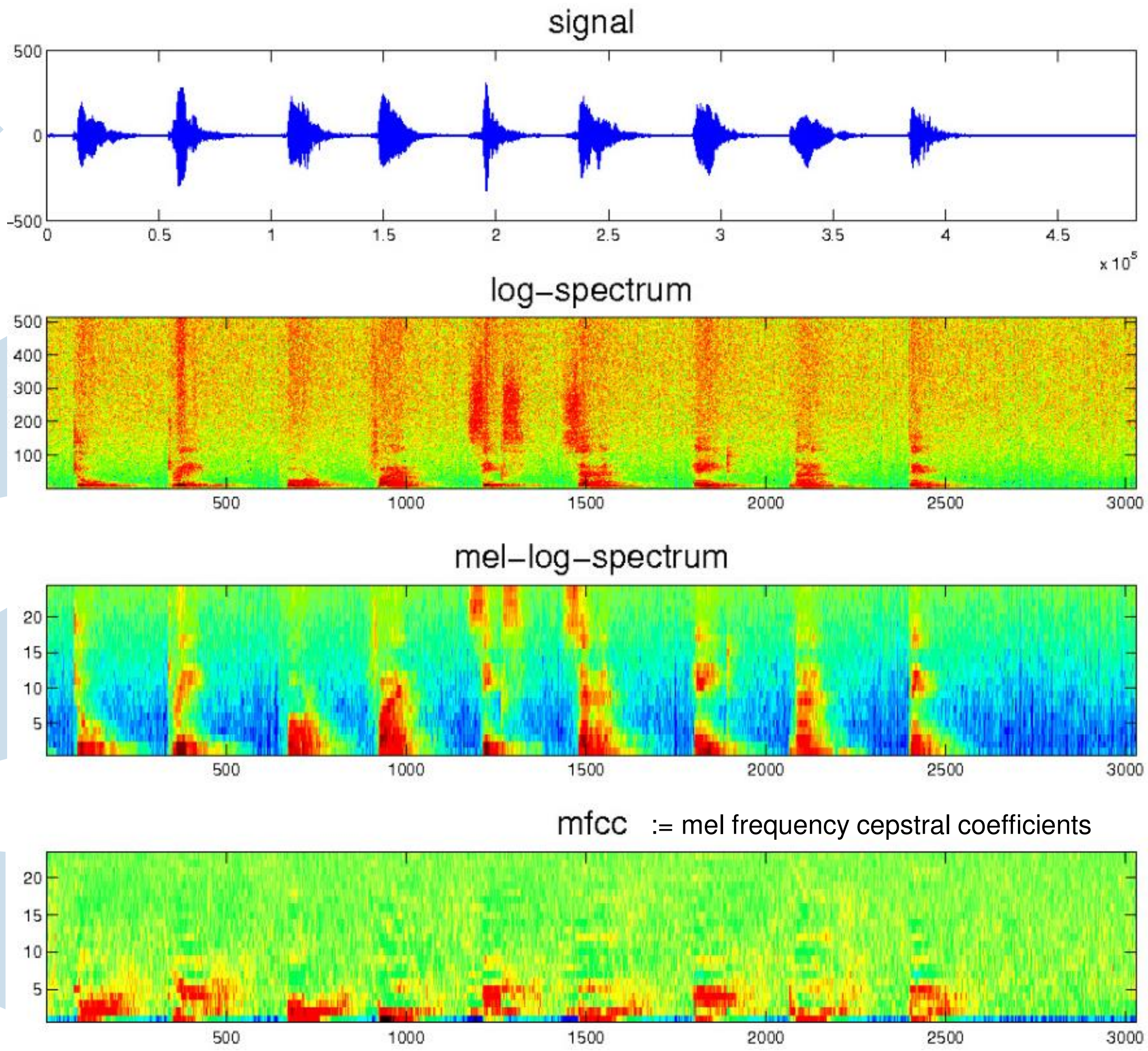


https://en.wikipedia.org/wiki/Decorrelation

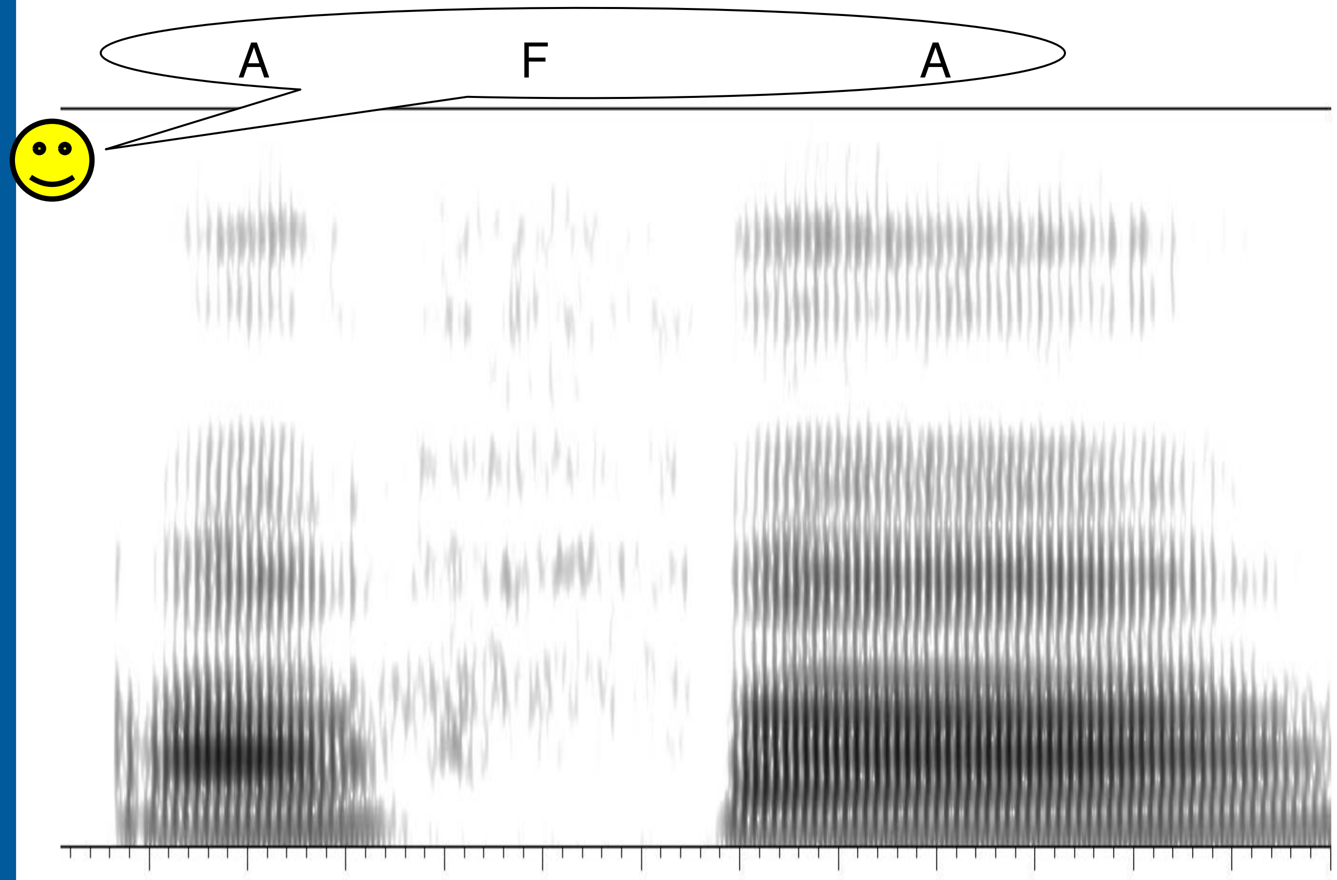
Feature Extraction for Speech Applications

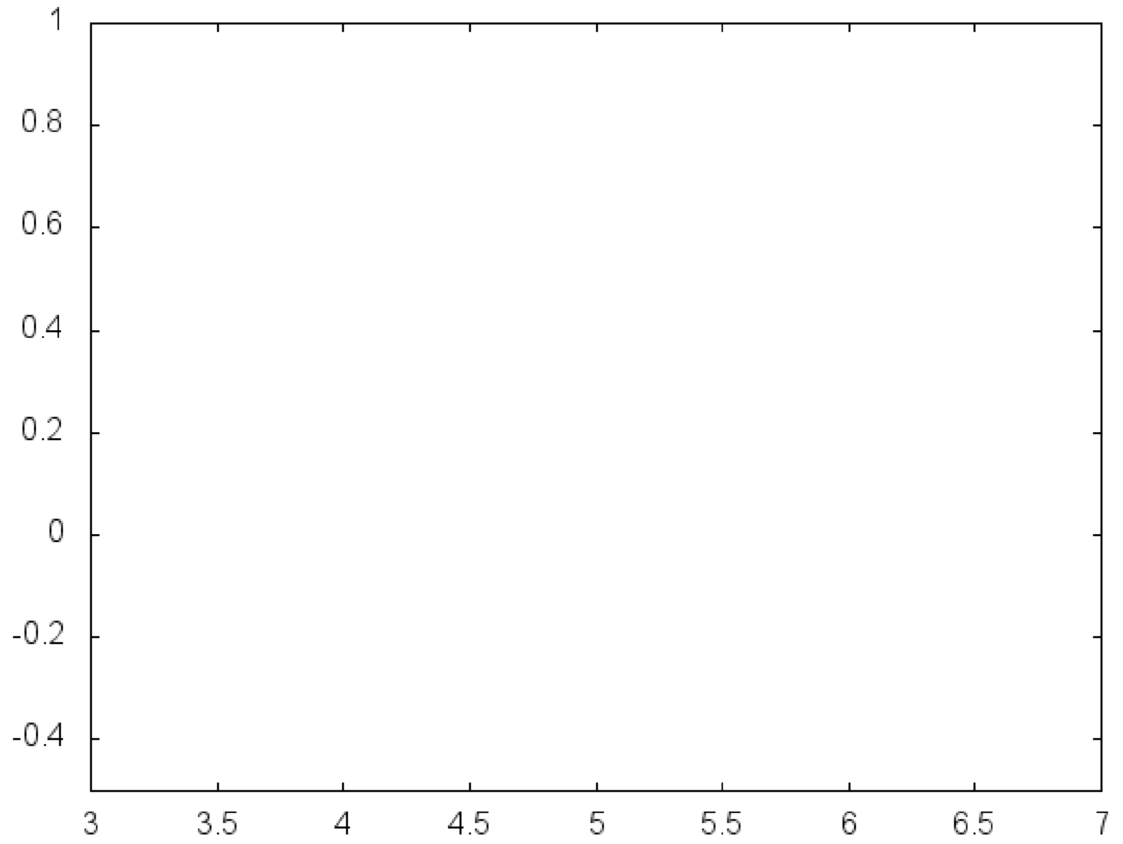
- N-Channel Audio Capturing
- High pass filter
- Pre-emphasis
- Windowing
- Fast Fourier Transformation
- Absolute Value
- Mel-Scale Filter bank
- Log-scale
- Discrete Cosine Transformation II

Speech/Speaker Recognition

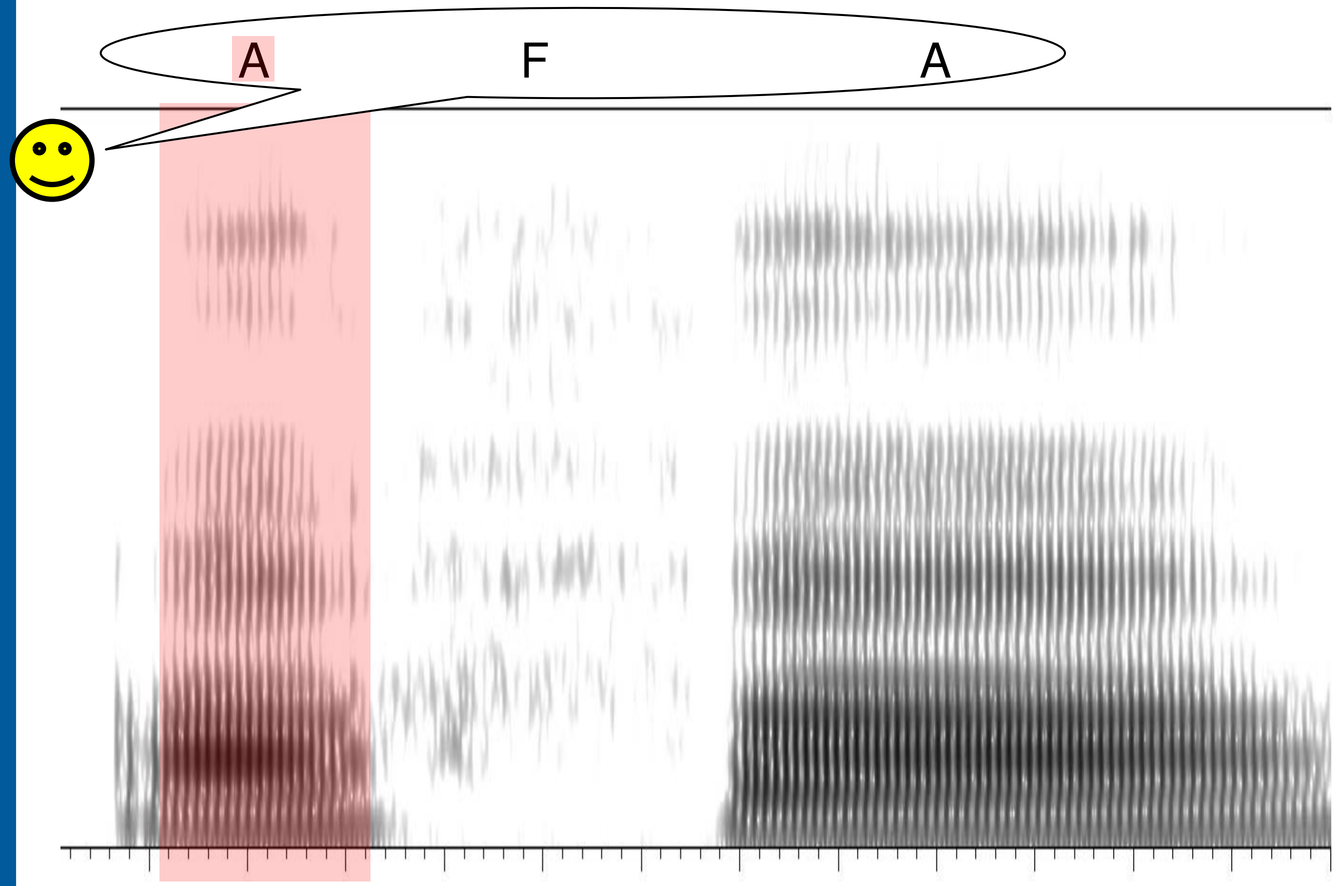


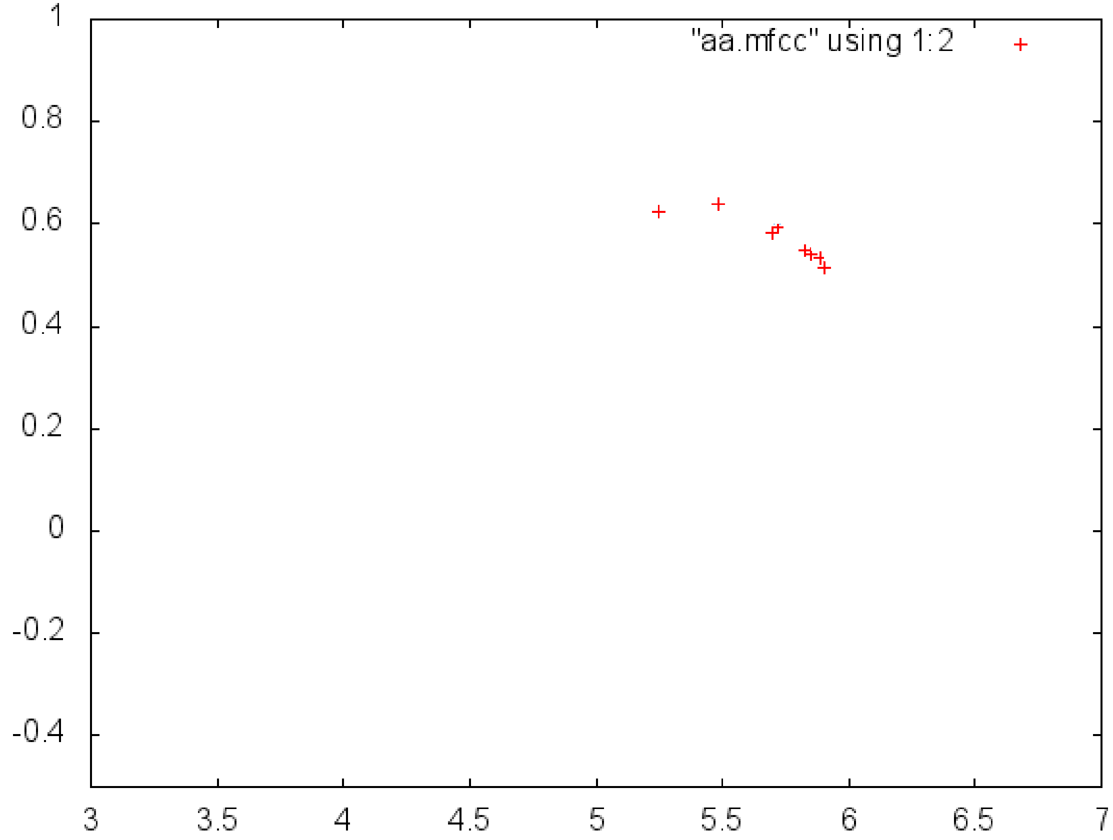
MFCC features are use for Automatic Speech Recognition



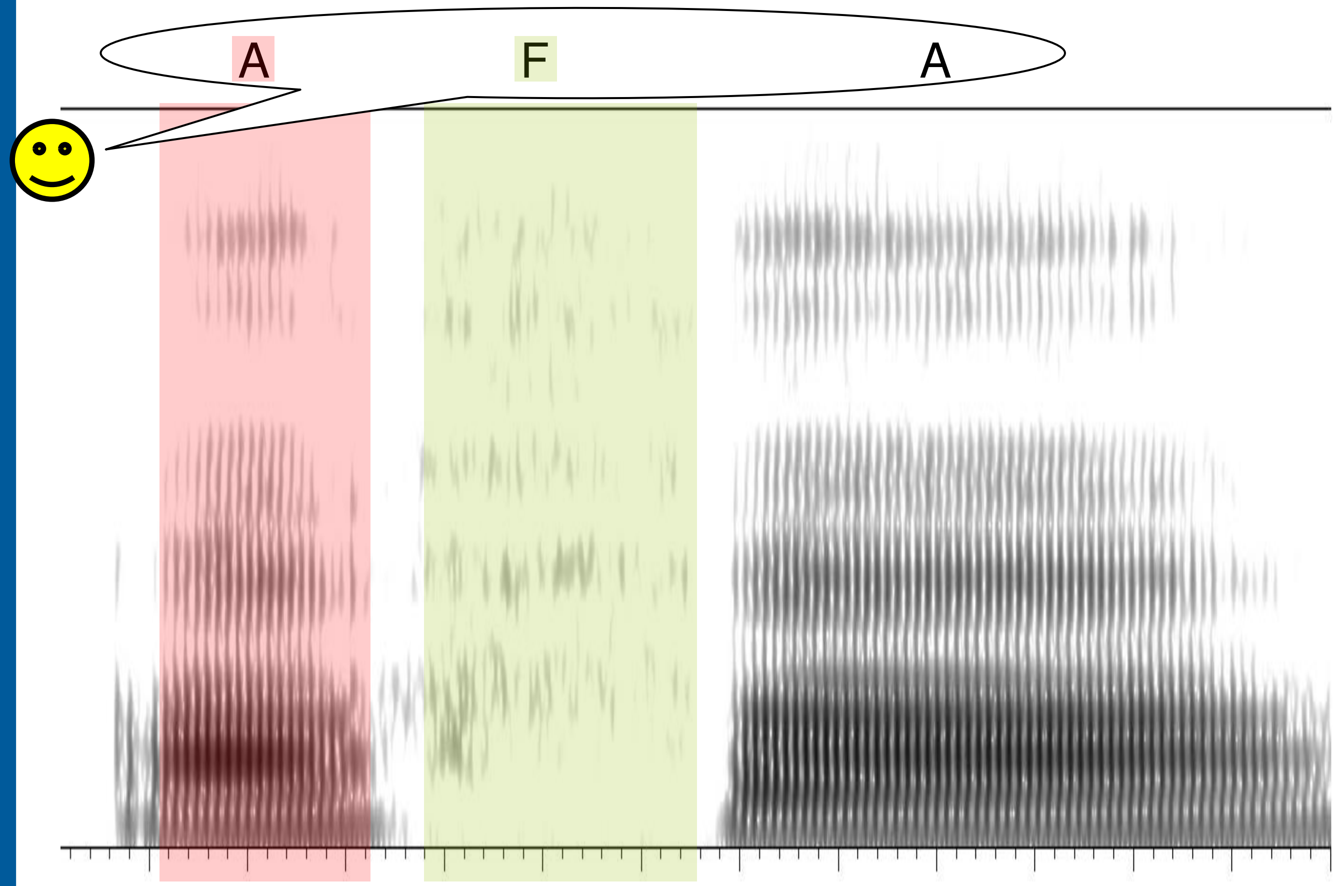


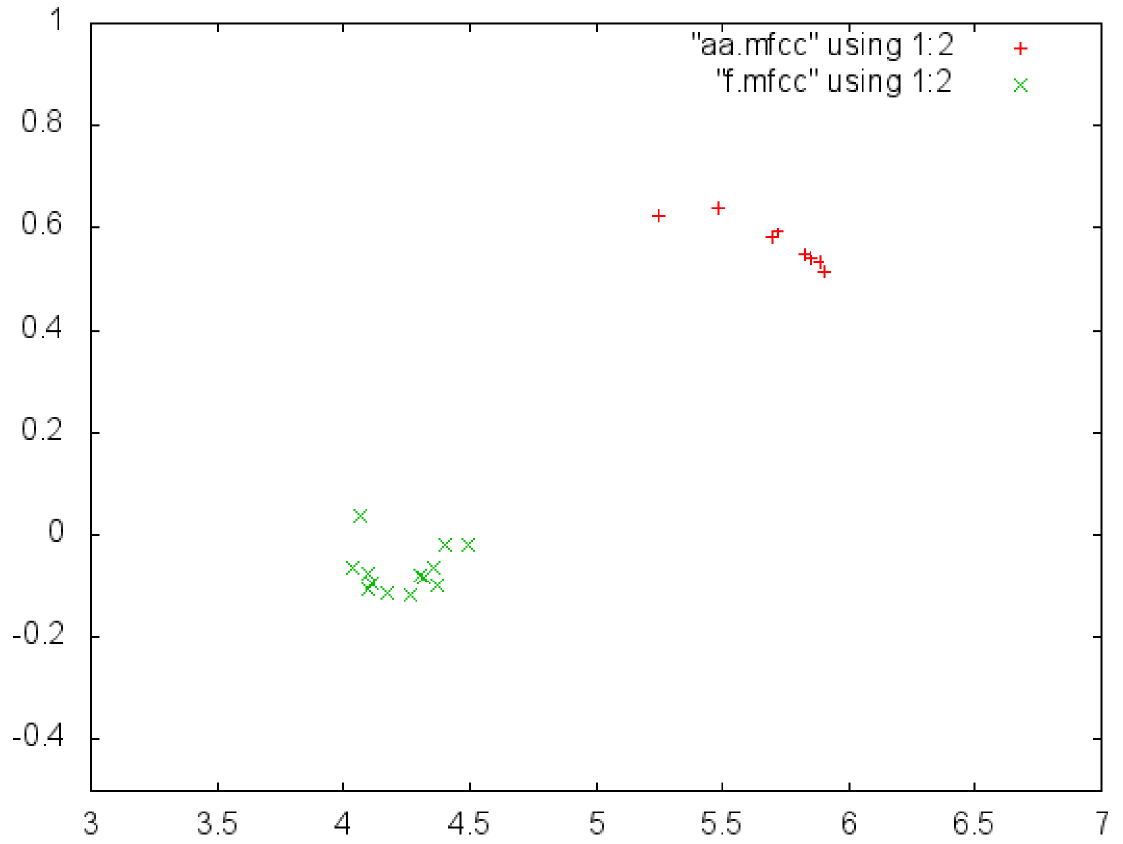
MFCC features are use for Automatic Speech Recognition



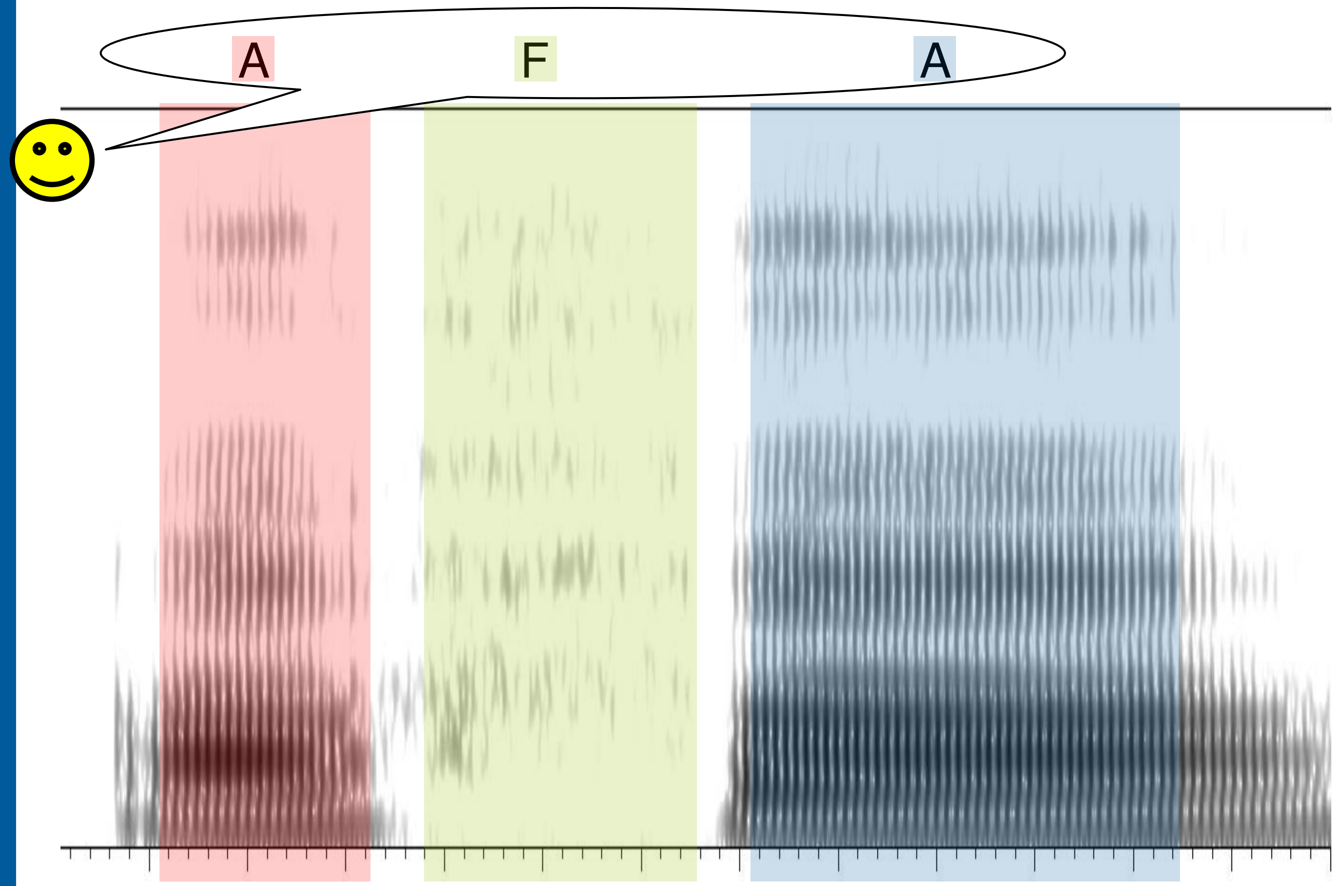


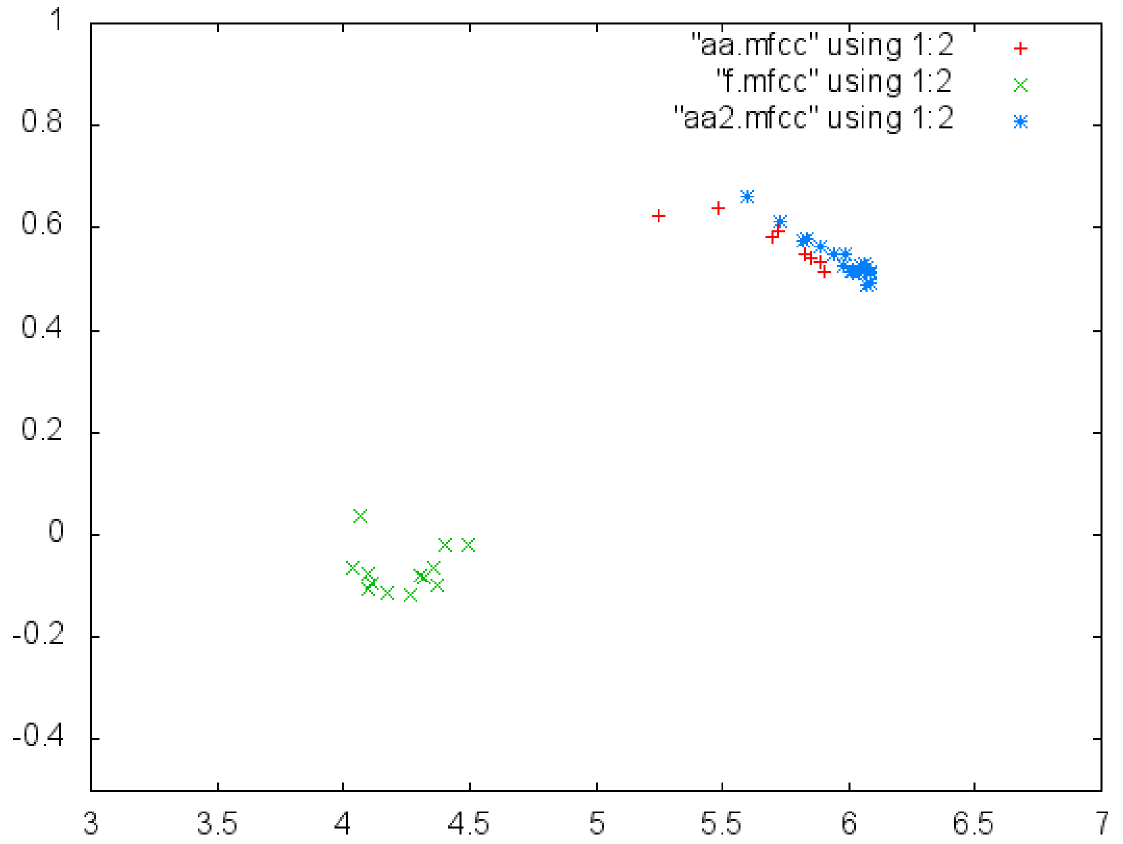
MFCC features are use for Automatic Speech Recognition



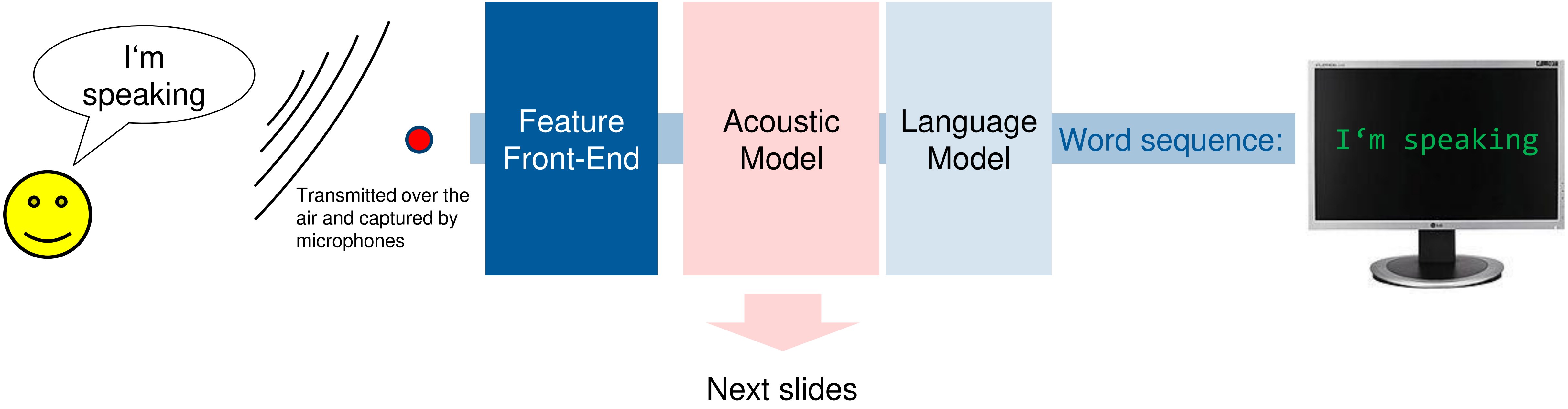


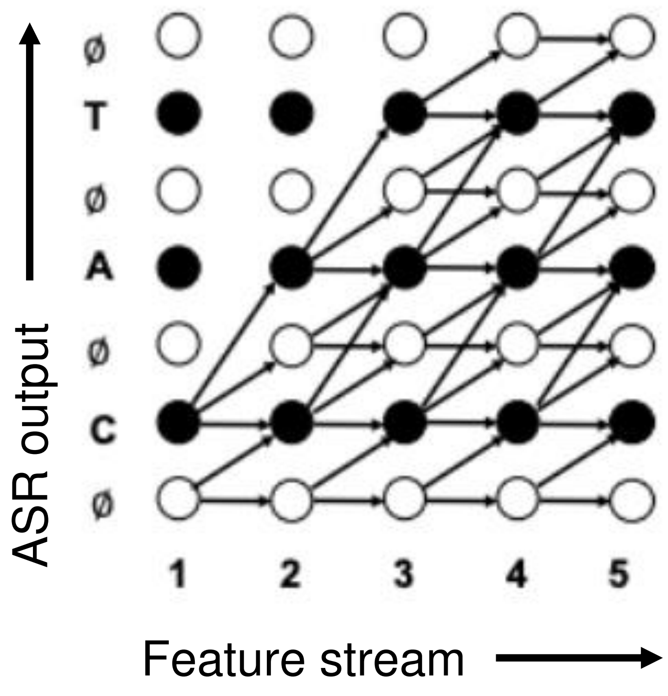
MFCC features are use for Automatic Speech Recognition





Automatic Speech Recognition, is the technology that allows human beings to use their voices to speak with a computer interface in a way that, in its most sophisticated variations, resembles normal human conversation.

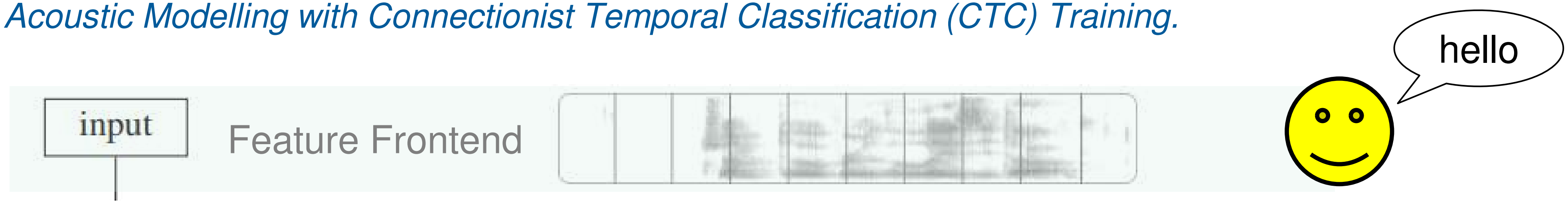




ASR output
Ø
T
Ø
A
Ø
C
Ø
1
2
3
4
5
Feature stream

Automatic Speech Recognition

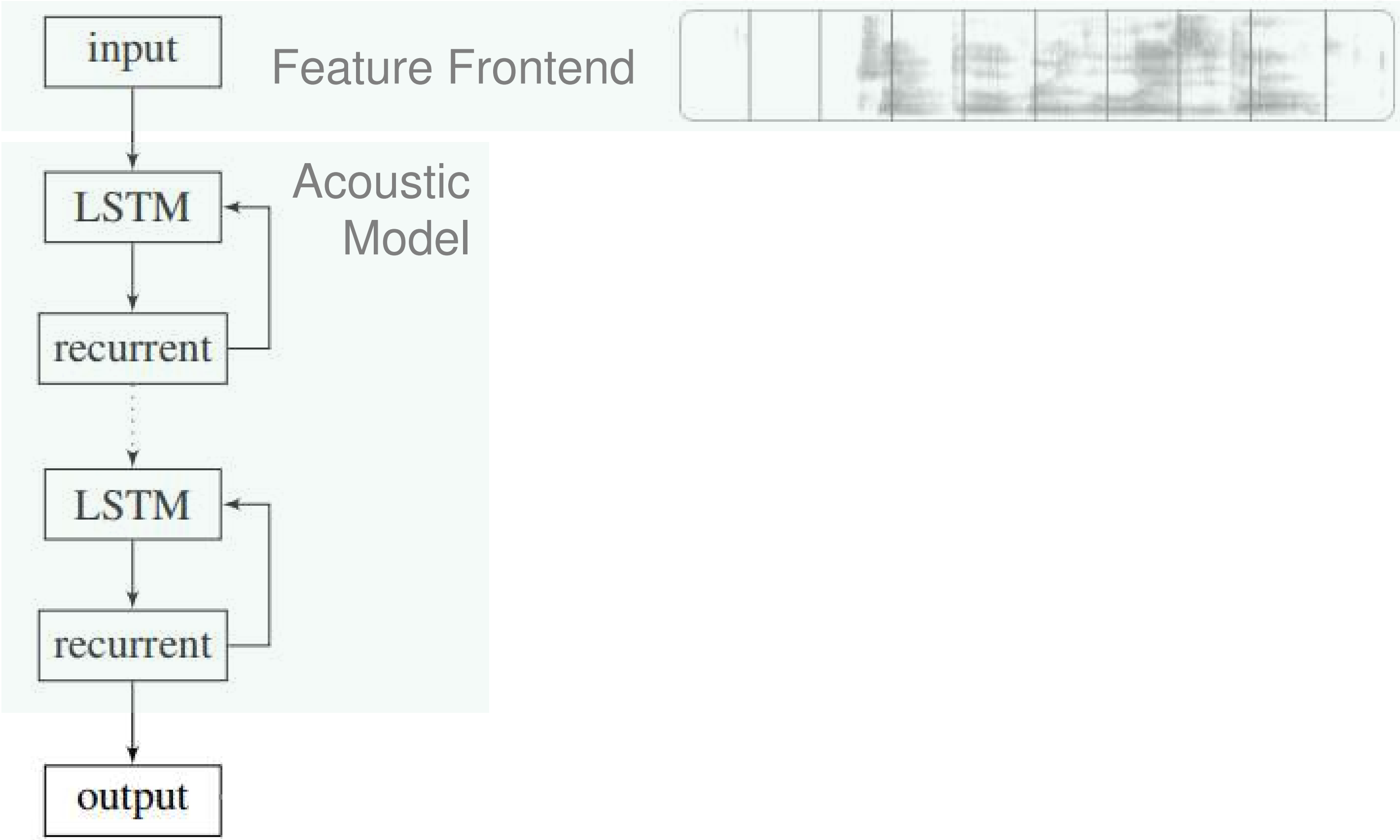
Acoustic Modelling with Connectionist Temporal Classification (CTC) Training.

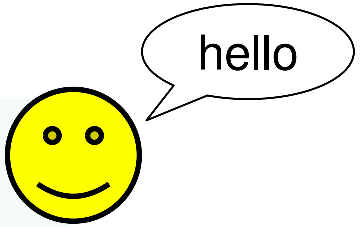


Best Overview: https://distill.pub/2017/ctc/

Automatic Speech Recognition

Acoustic Modelling with Connectionist Temporal Classification (CTC) Training.

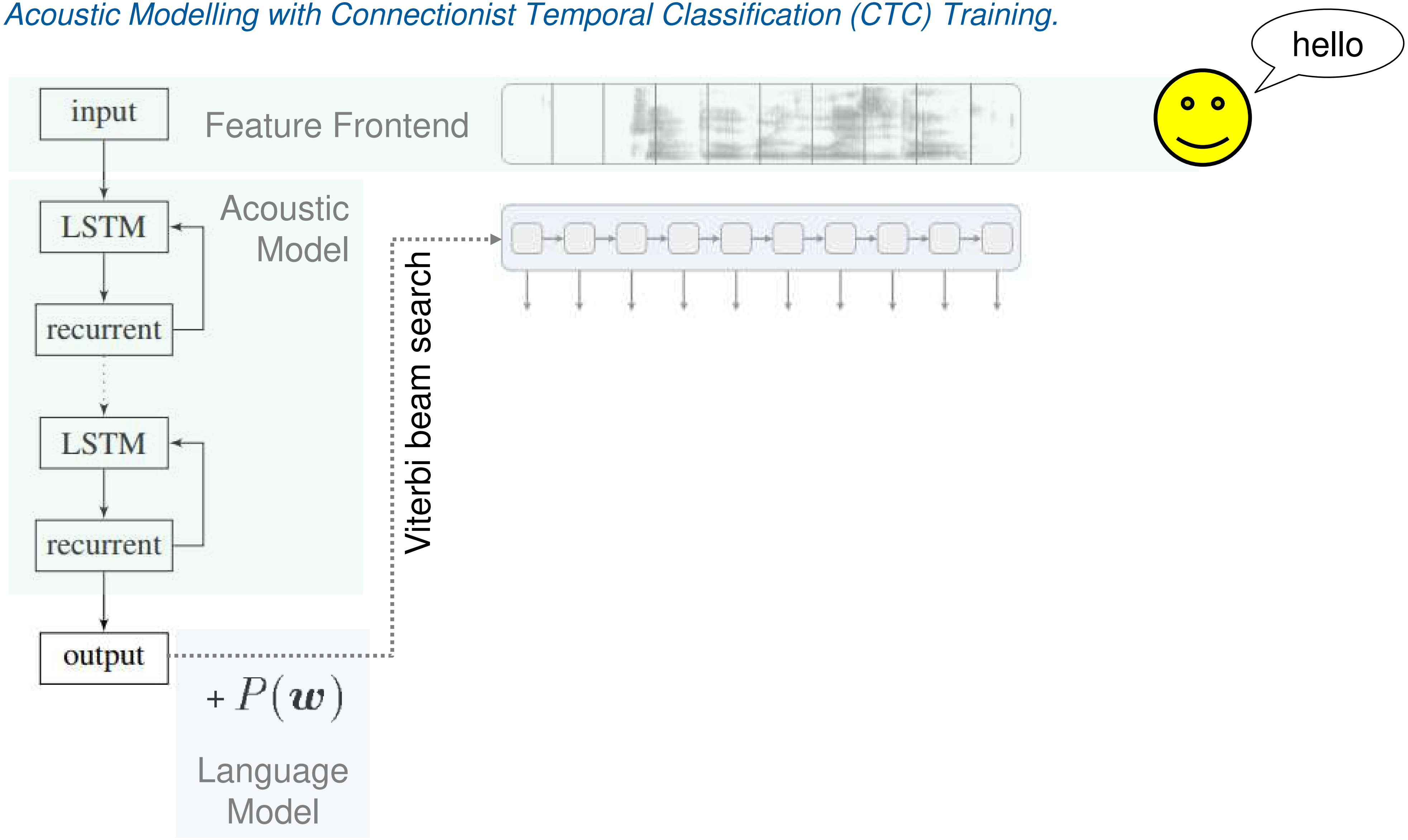




Best Overview: https://distill.pub/2017/ctc/

Automatic Speech Recognition

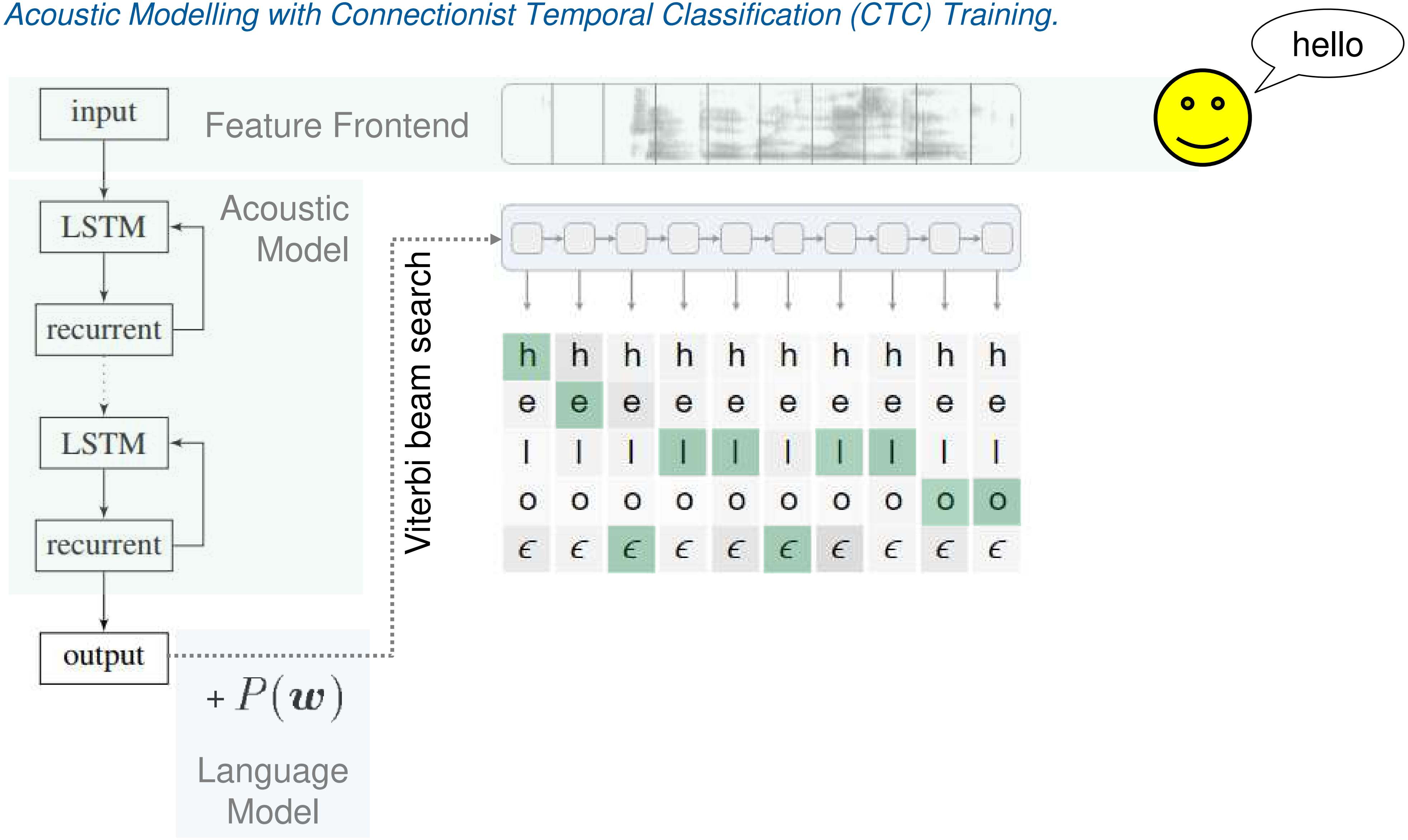
Acoustic Modelling with Connectionist Temporal Classification (CTC) Training.



Best Overview: https://distill.pub/2017/ctc/

Automatic Speech Recognition

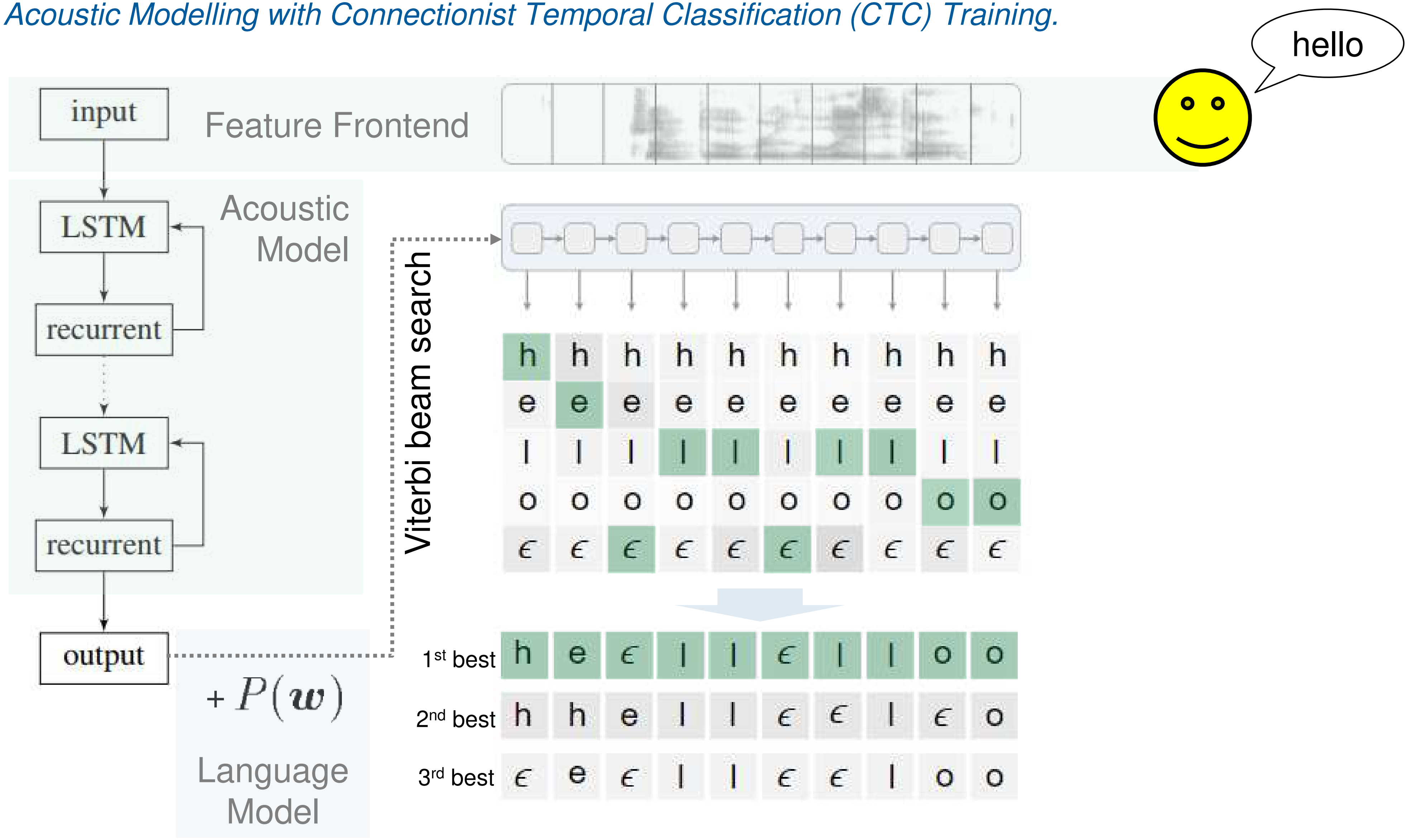
Acoustic Modelling with Connectionist Temporal Classification (CTC) Training.



Best Overview: https://distill.pub/2017/ctc/

Automatic Speech Recognition

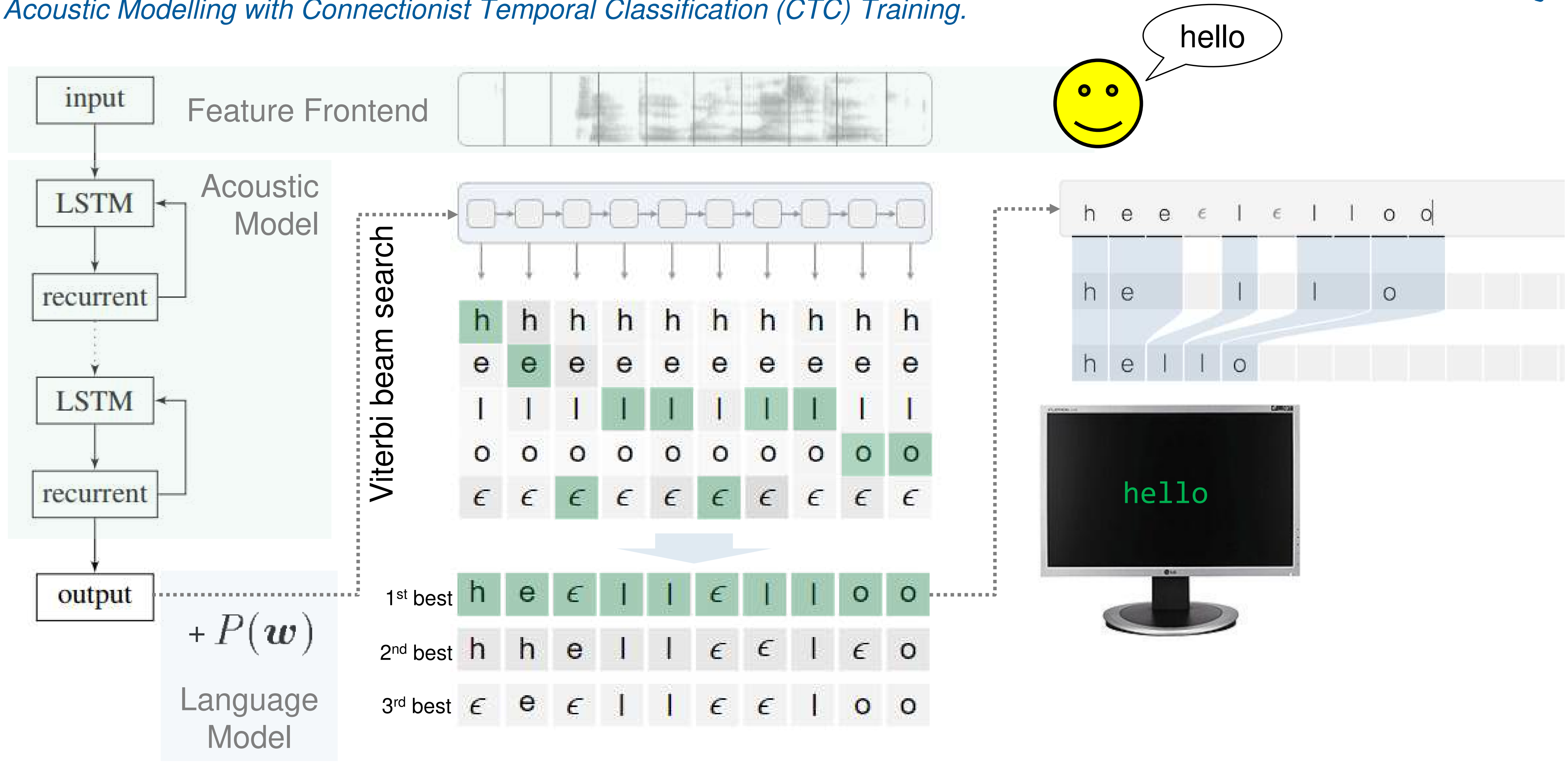
Acoustic Modelling with Connectionist Temporal Classification (CTC) Training.



Best Overview: https://distill.pub/2017/ctc/

Automatic Speech Recognition

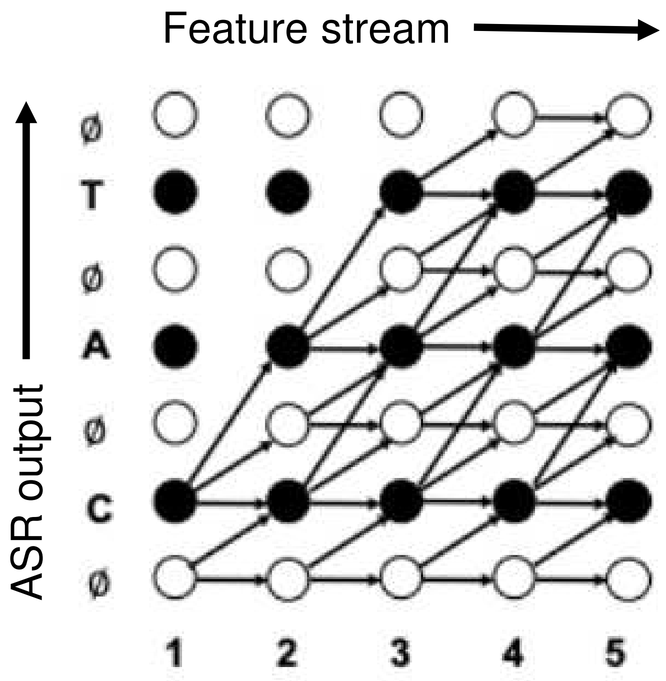
Acoustic Modelling with Connectionist Temporal Classification (CTC) Training.



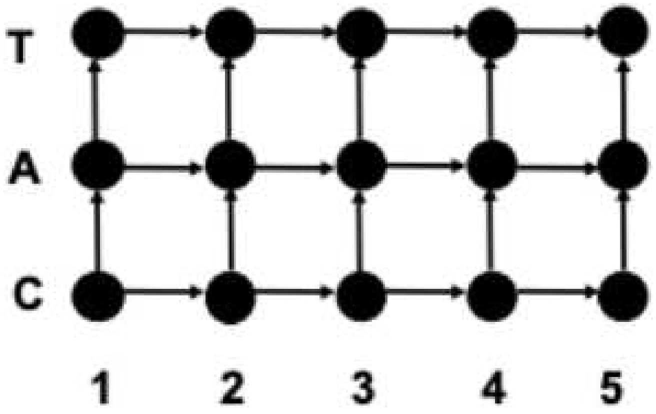
Best Overview: https://distill.pub/2017/ctc/

How is it all connected?

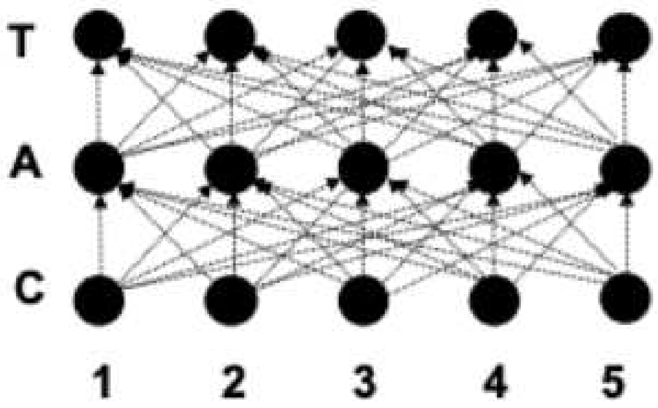
CTC vs. RNN-T vs. Attention



CTC

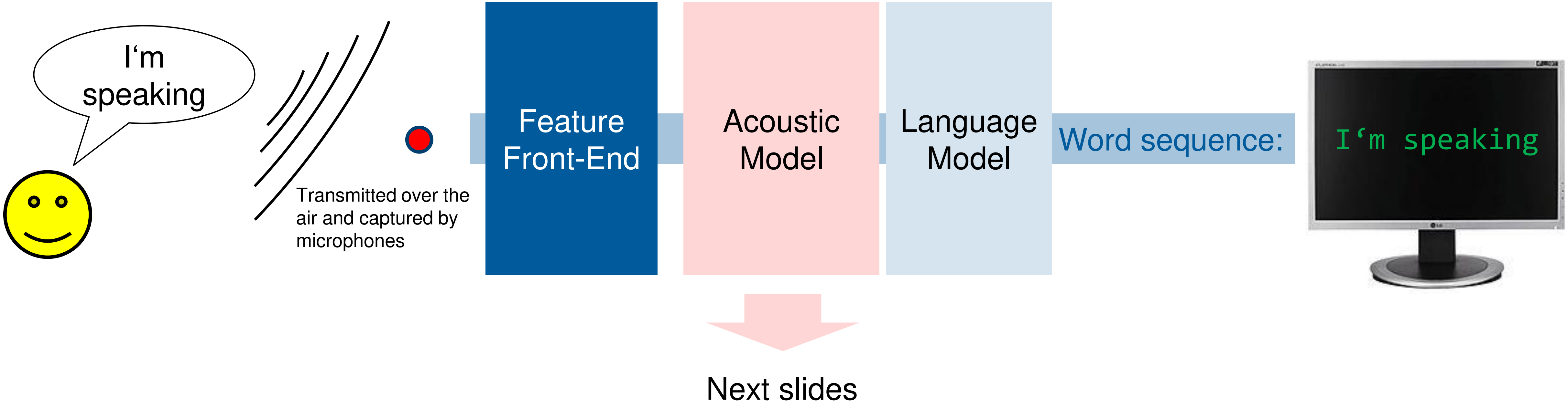


RNN-Transducer



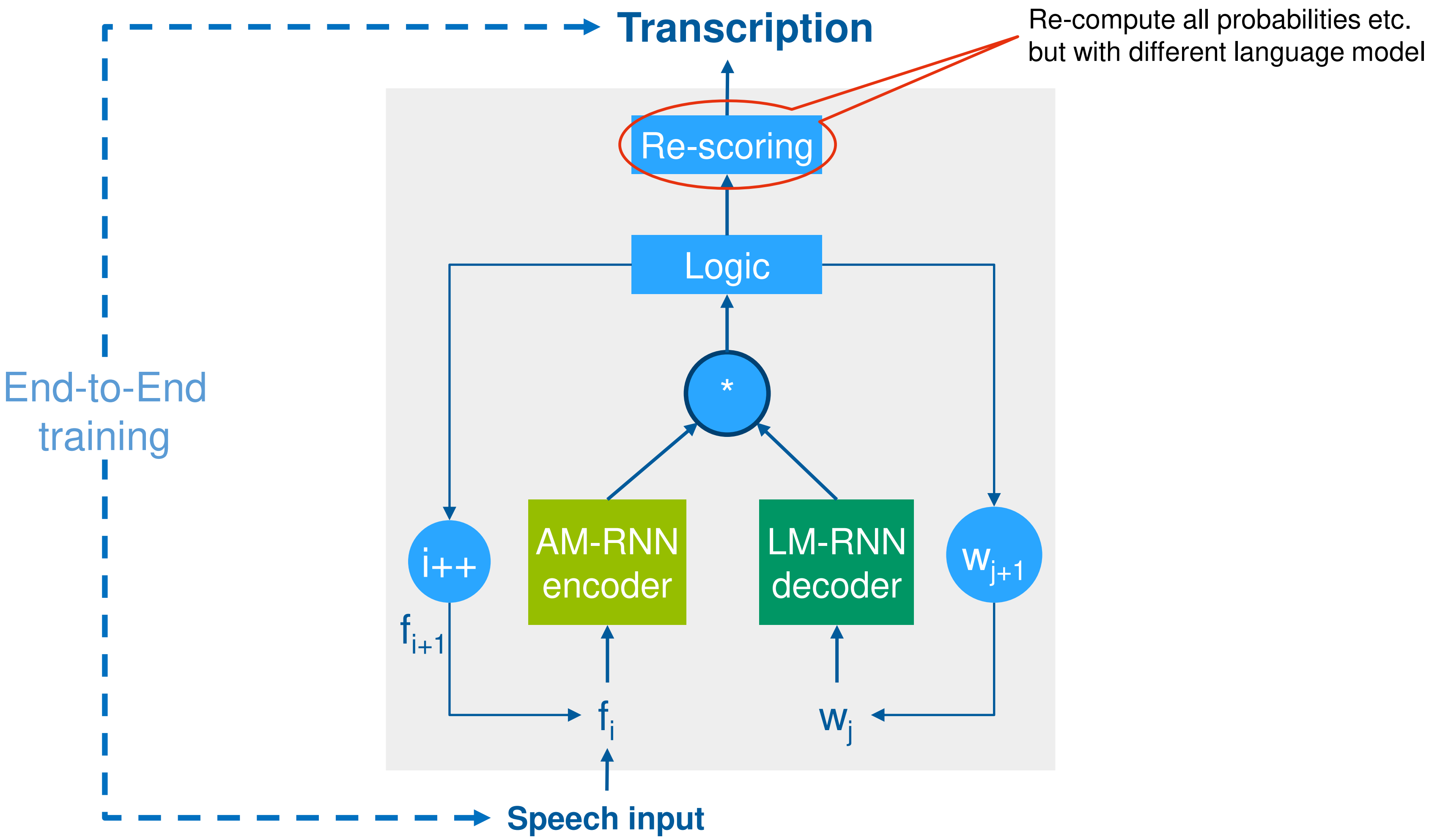
Attention

Automatic Speech Recognition, is the technology that allows human beings to use their voices to speak with a computer interface in a way that, in its most sophisticated variations, resembles normal human conversation.



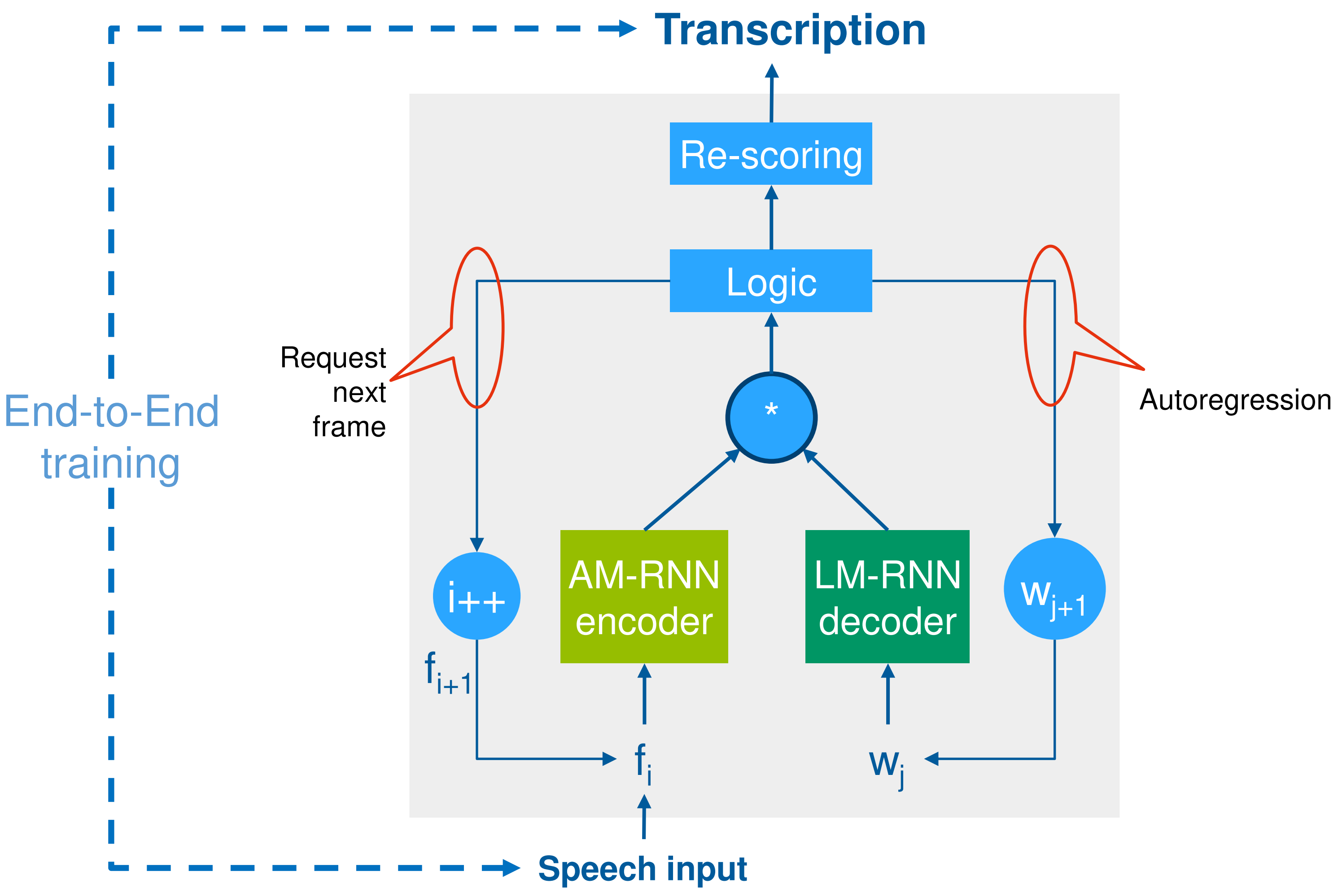
End-to-End Trained Automatic Speech Recognition

Recurrent Neural Network Transducer



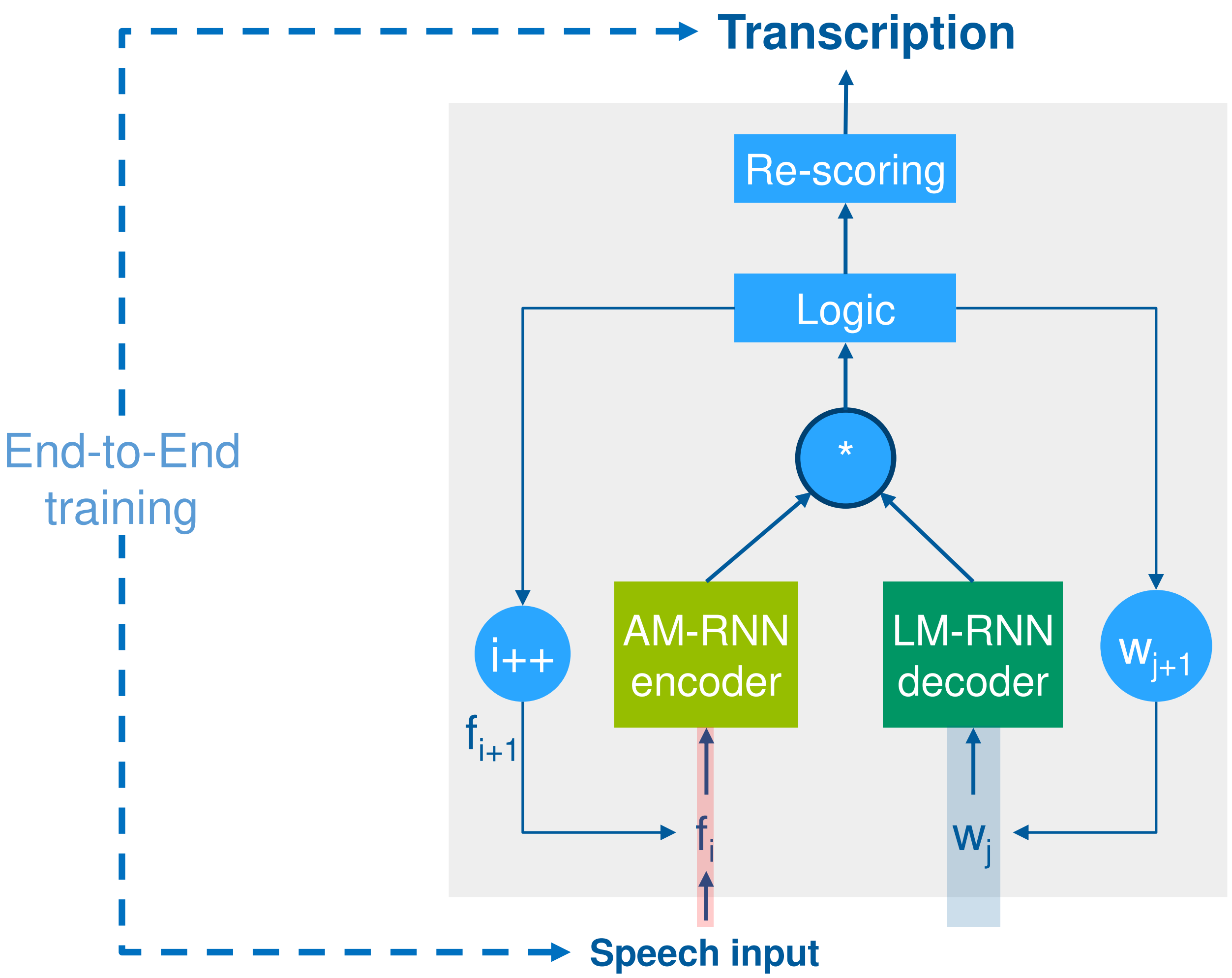
End-to-End Trained Automatic Speech Recognition

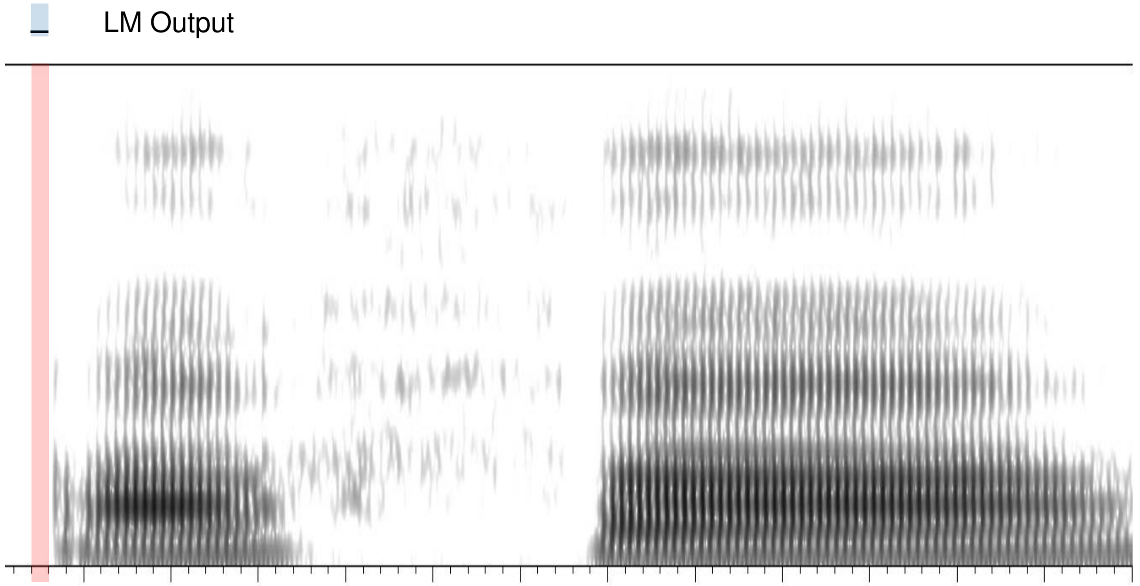
Recurrent Neural Network Transducer



End-to-End Trained Automatic Speech Recognition

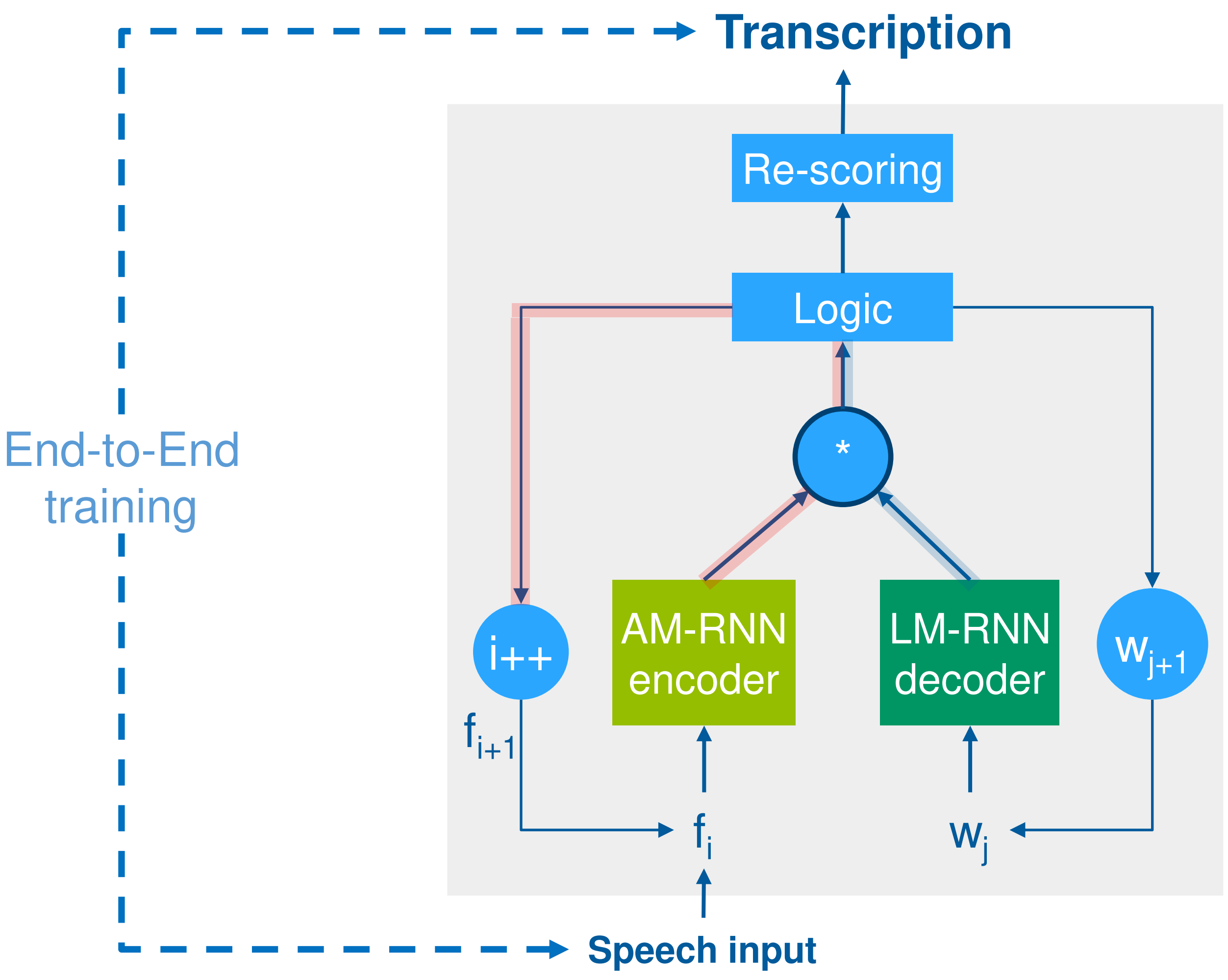
Recurrent Neural Network Transducer

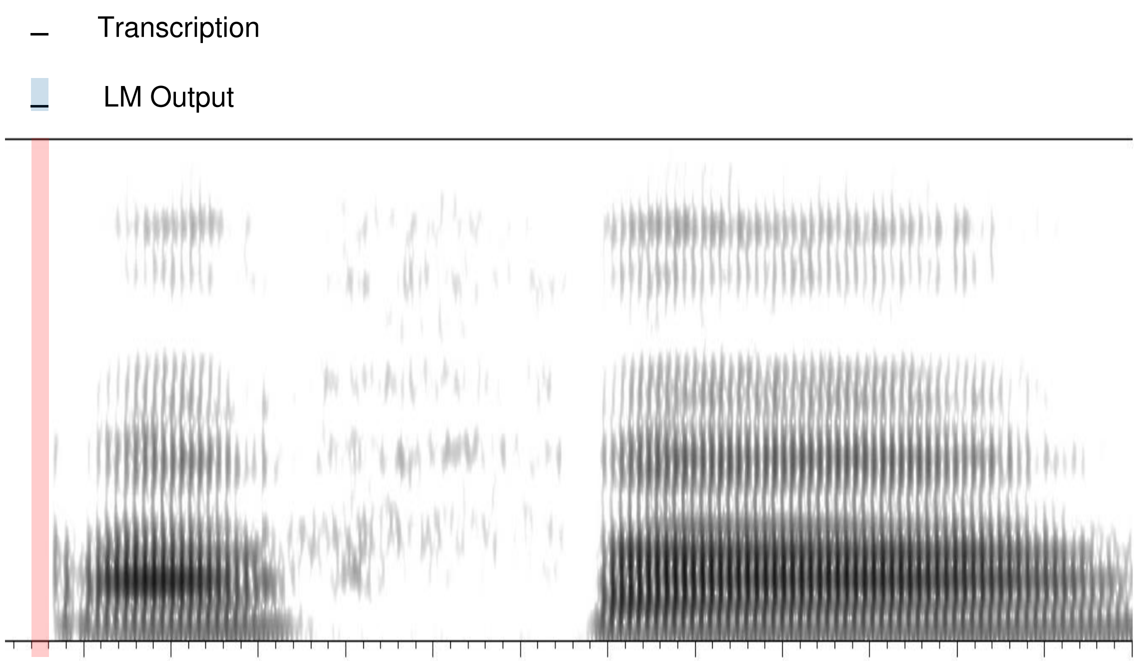




End-to-End Trained Automatic Speech Recognition

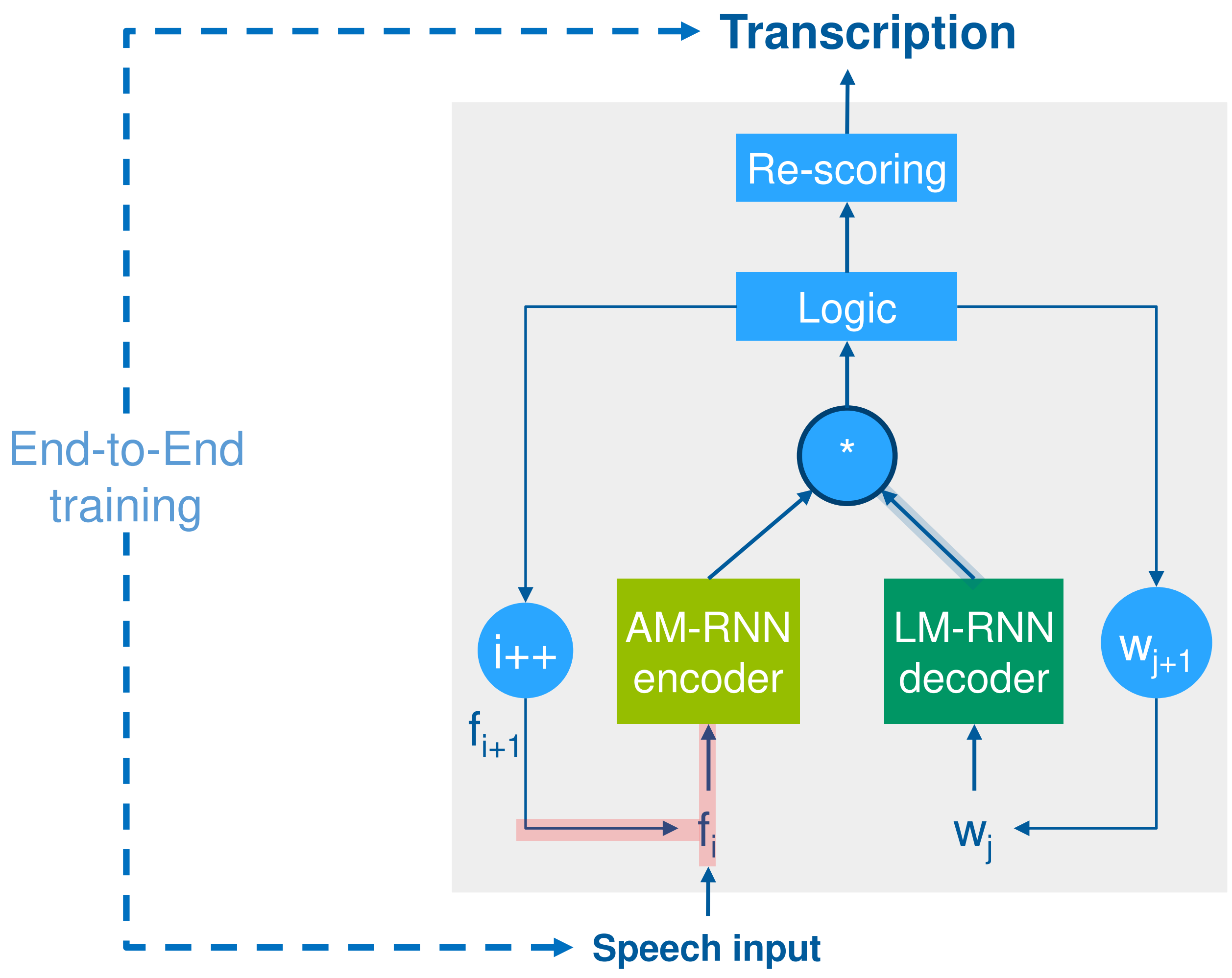
Recurrent Neural Network Transducer

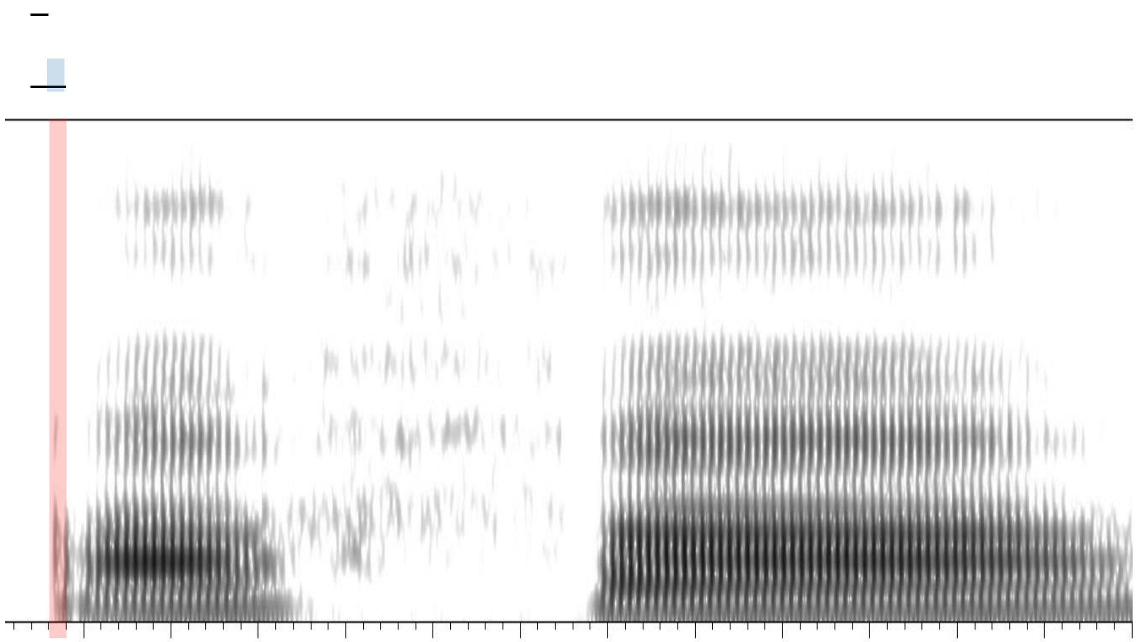




End-to-End Trained Automatic Speech Recognition

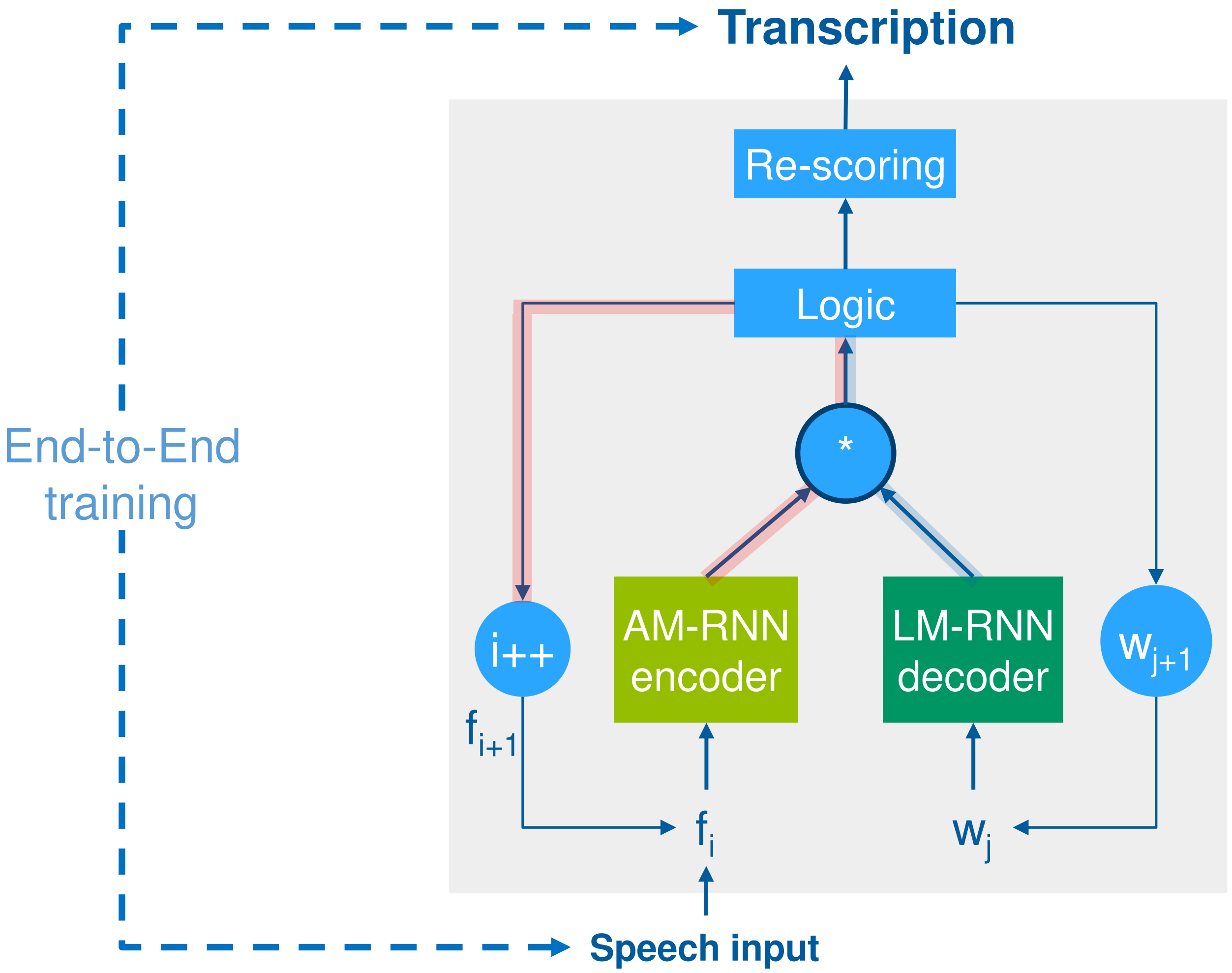
Recurrent Neural Network Transducer

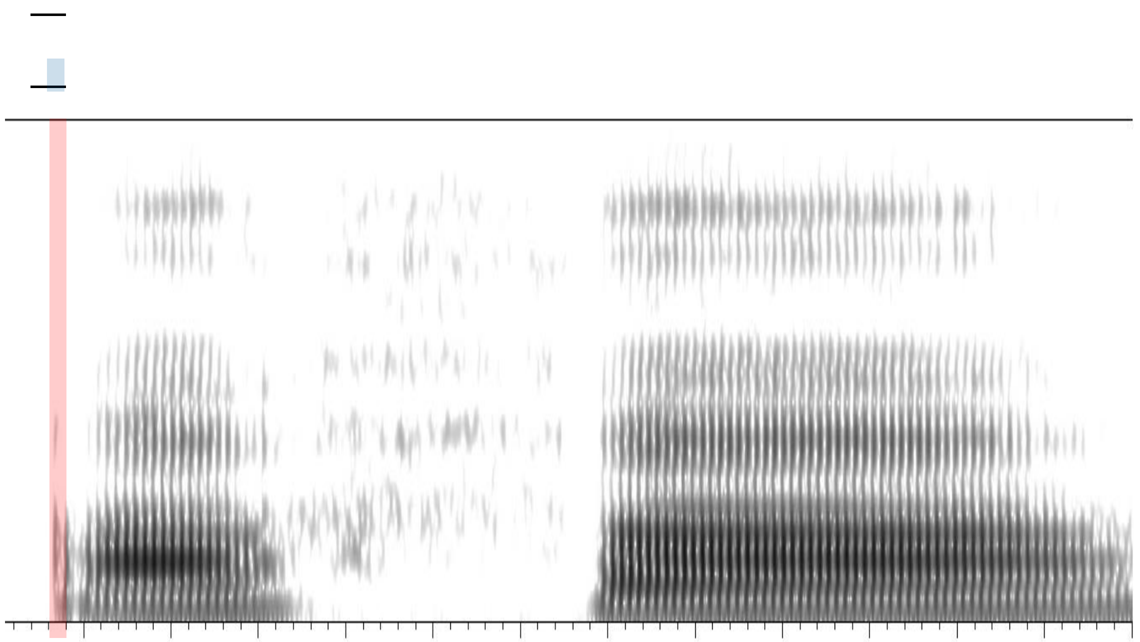




End-to-End Trained Automatic Speech Recognition

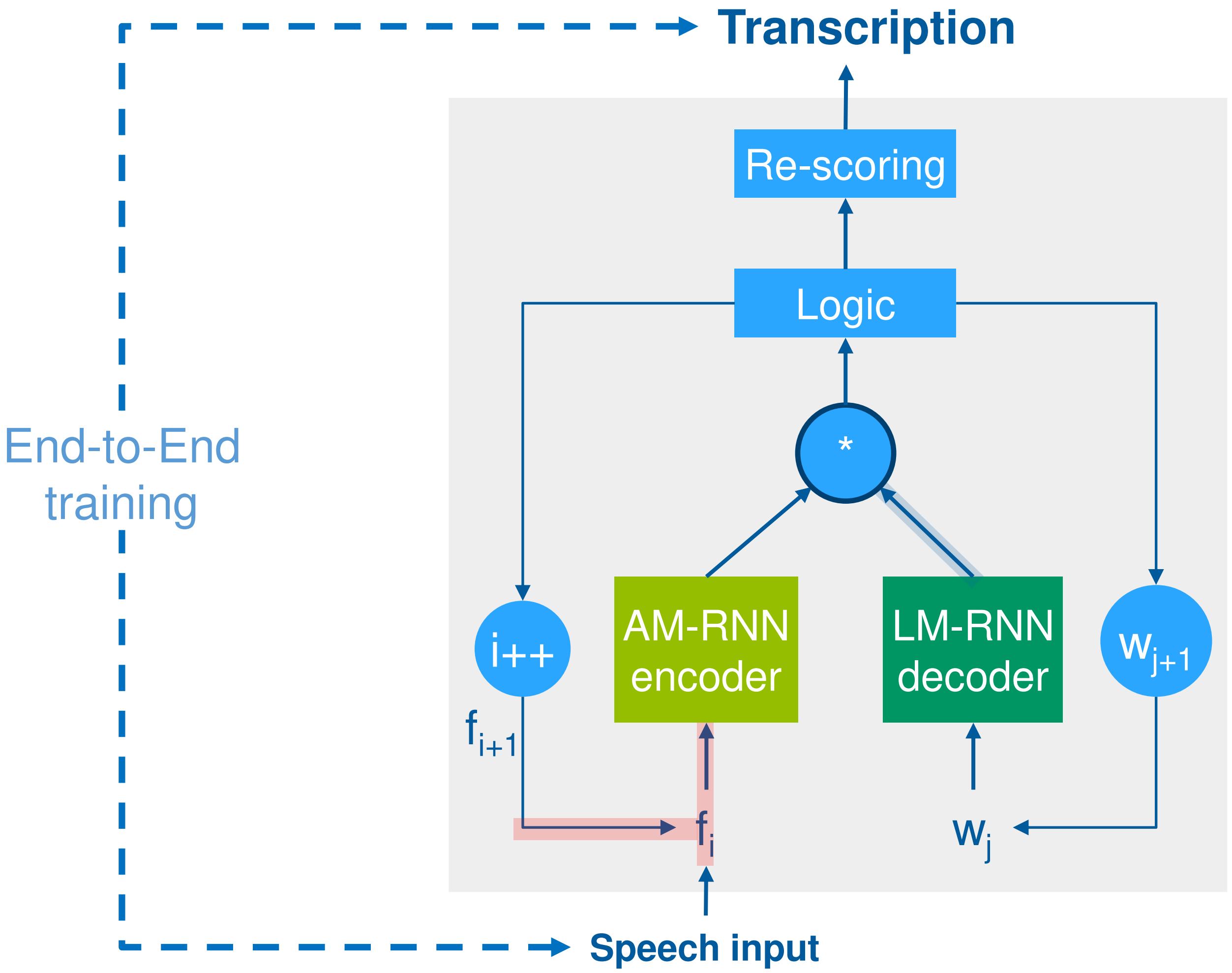
Recurrent Neural Network Transducer

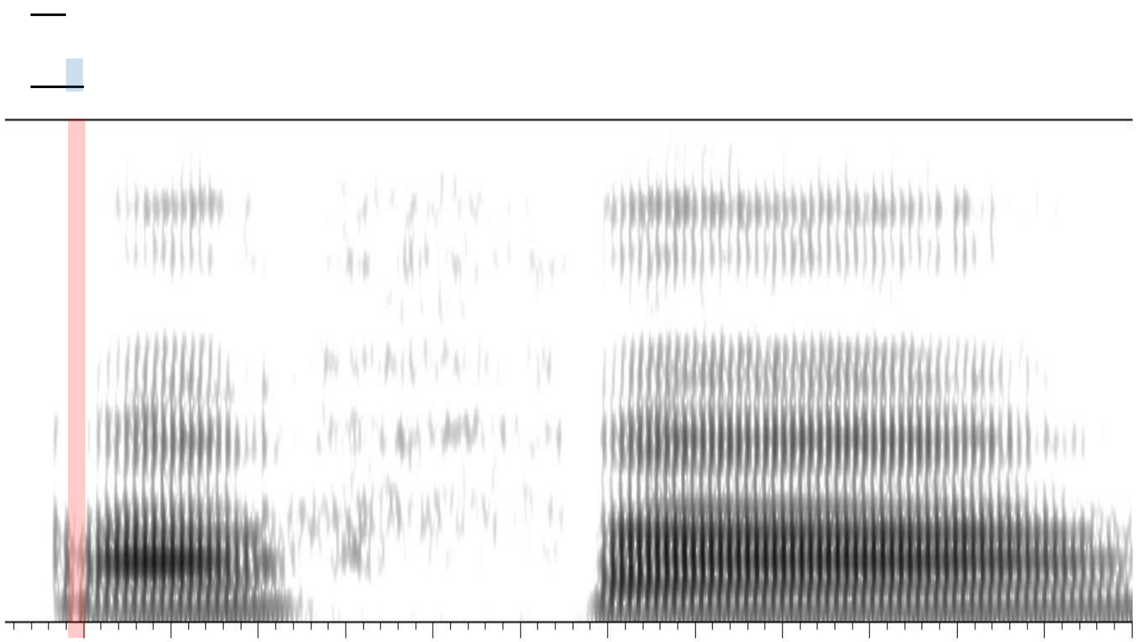




End-to-End Trained Automatic Speech Recognition

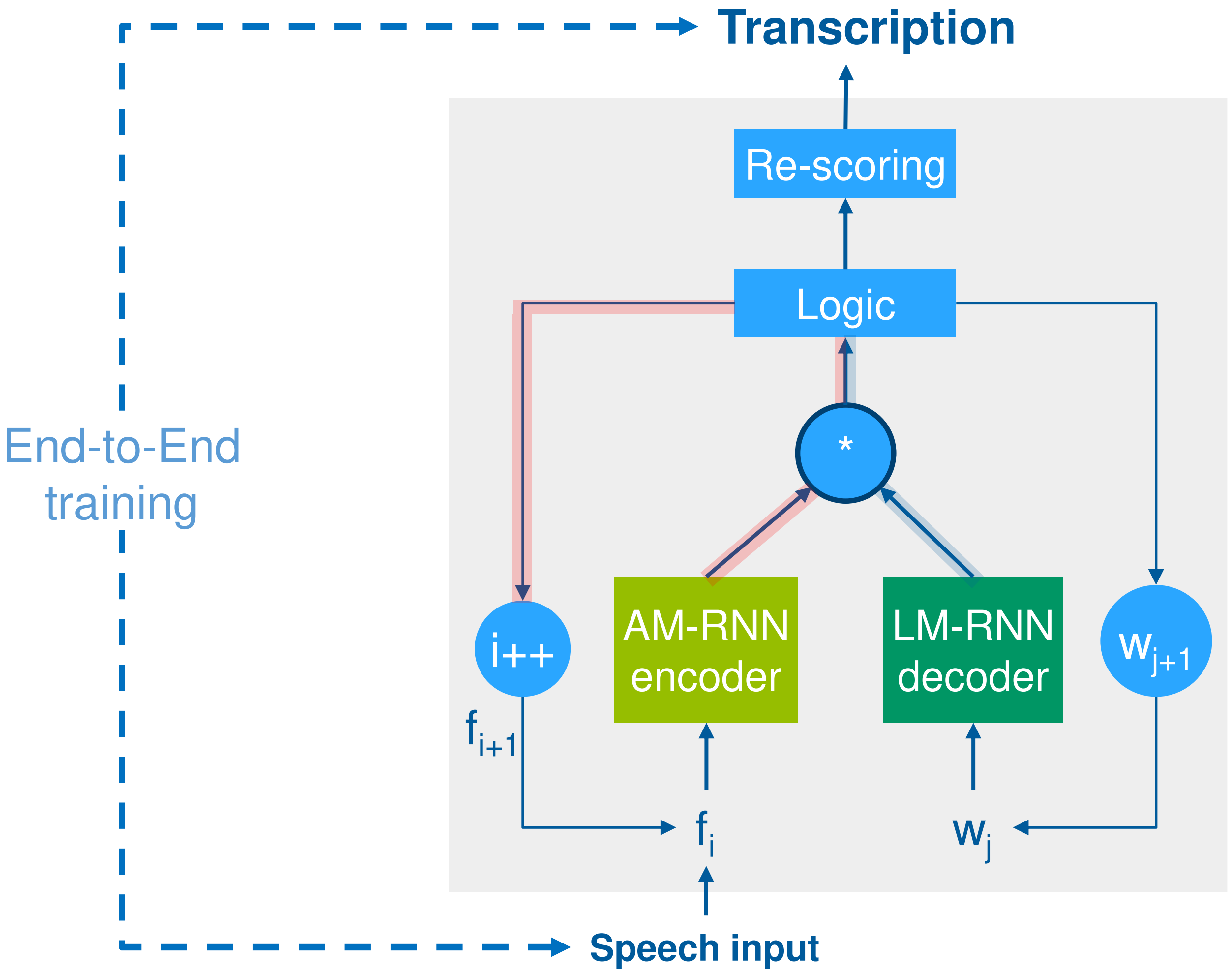
Recurrent Neural Network Transducer

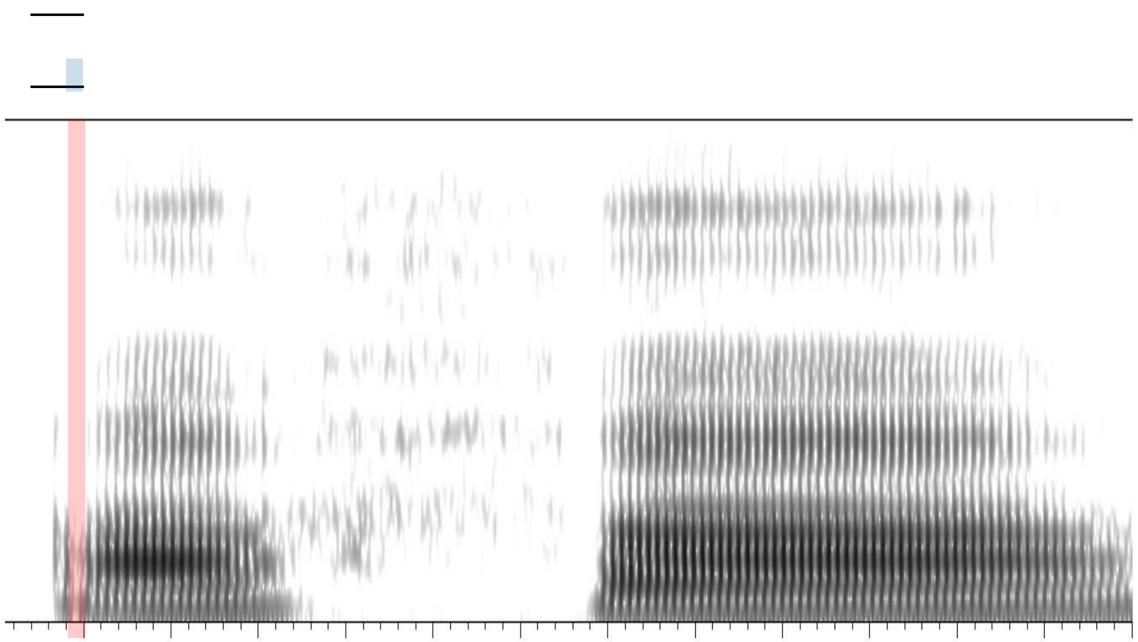




End-to-End Trained Automatic Speech Recognition

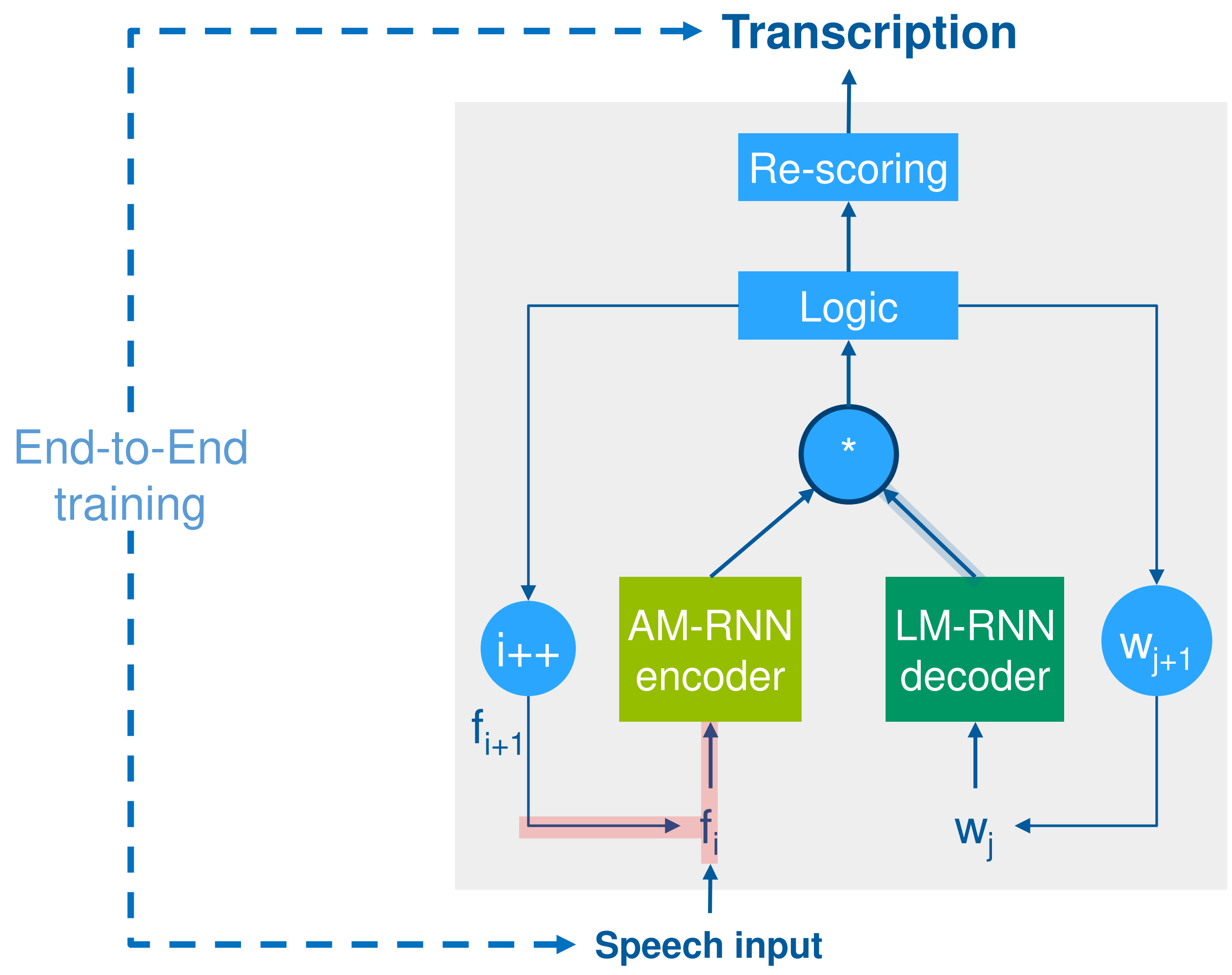
Recurrent Neural Network Transducer

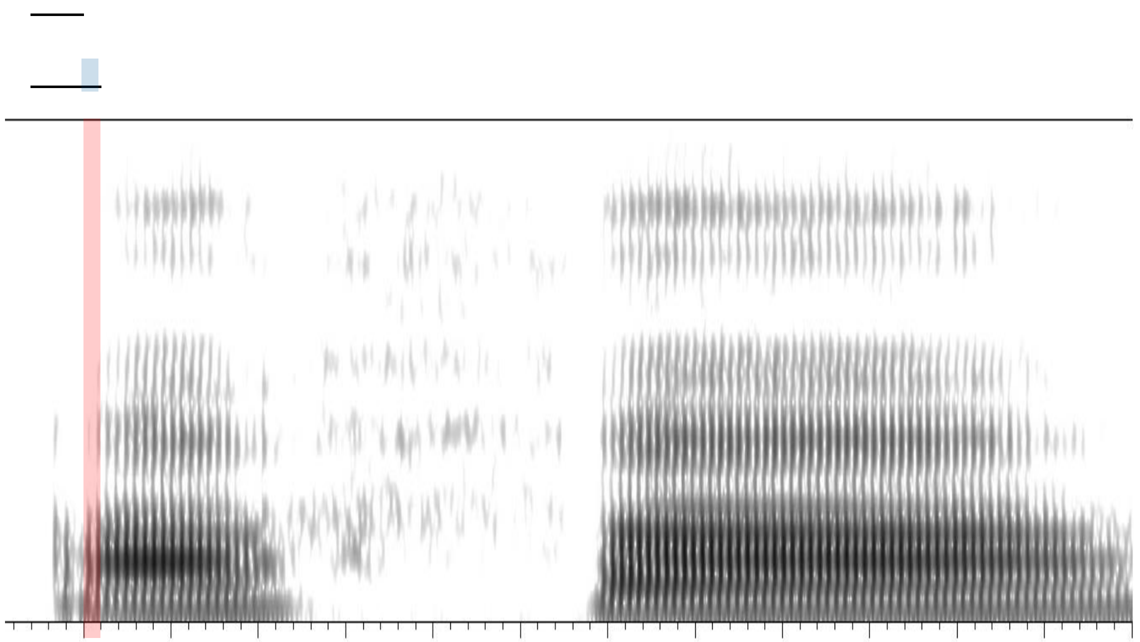




End-to-End Trained Automatic Speech Recognition

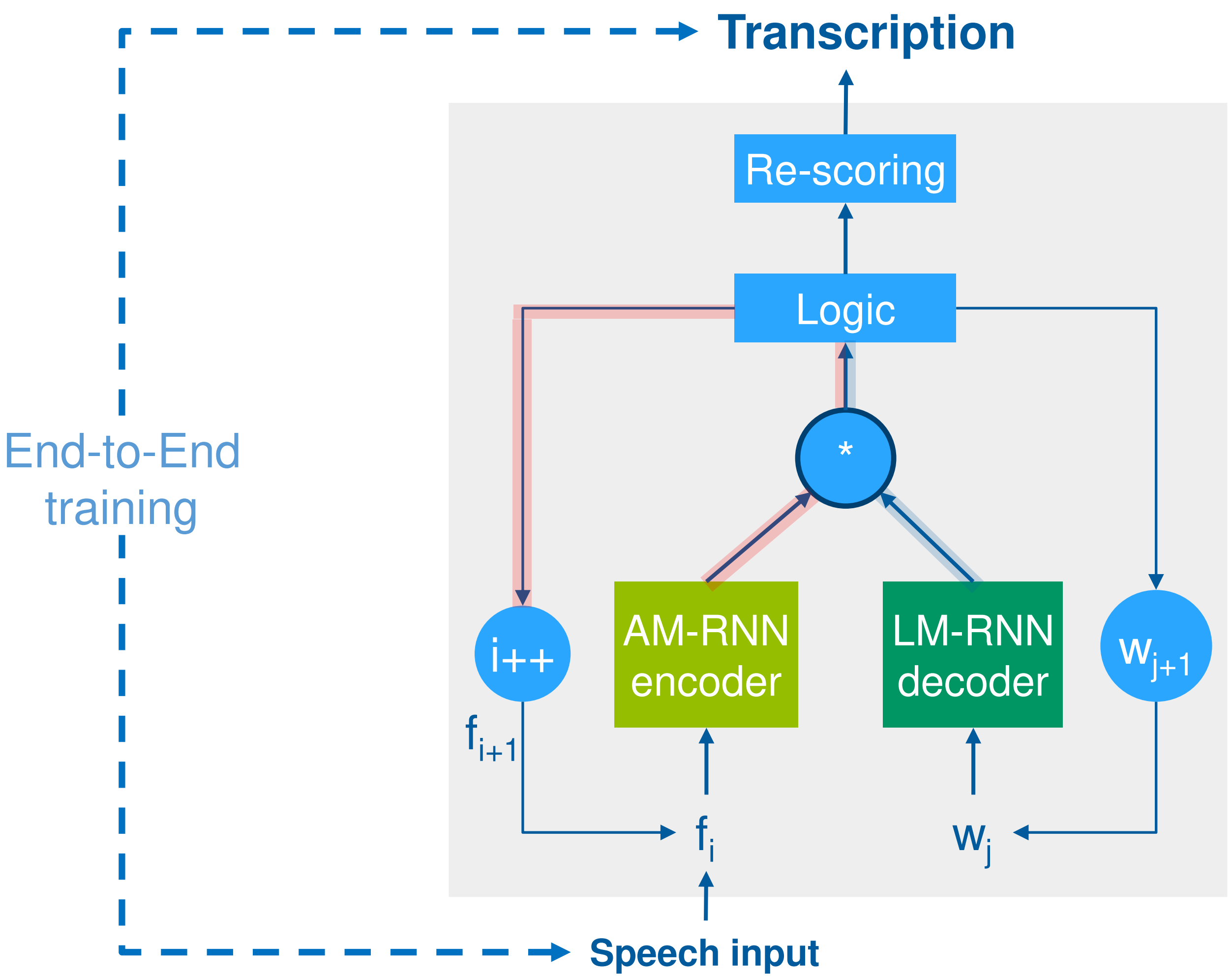
Recurrent Neural Network Transducer

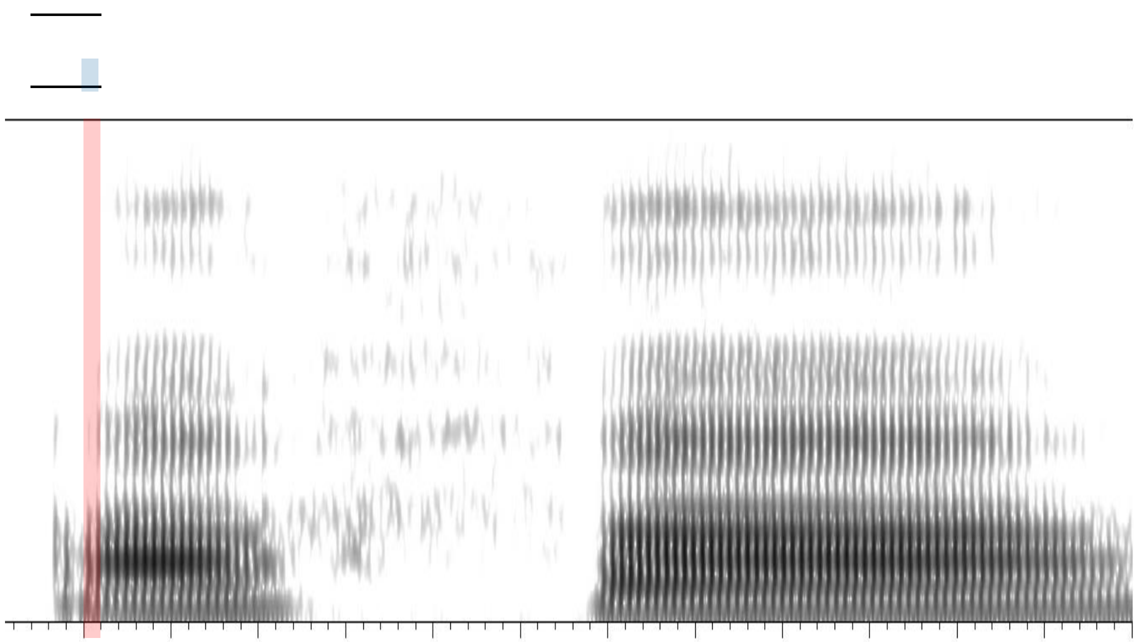




End-to-End Trained Automatic Speech Recognition

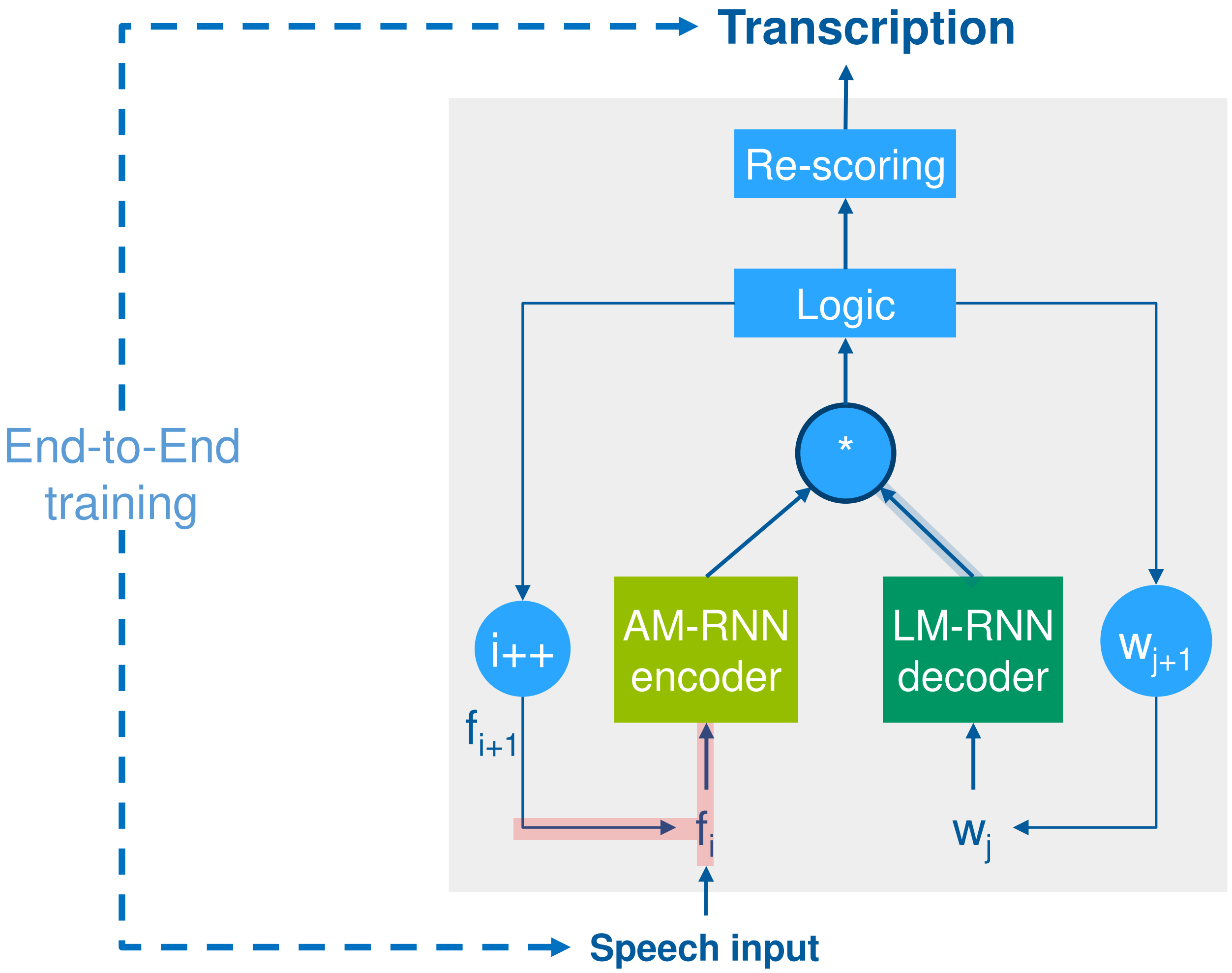
Recurrent Neural Network Transducer

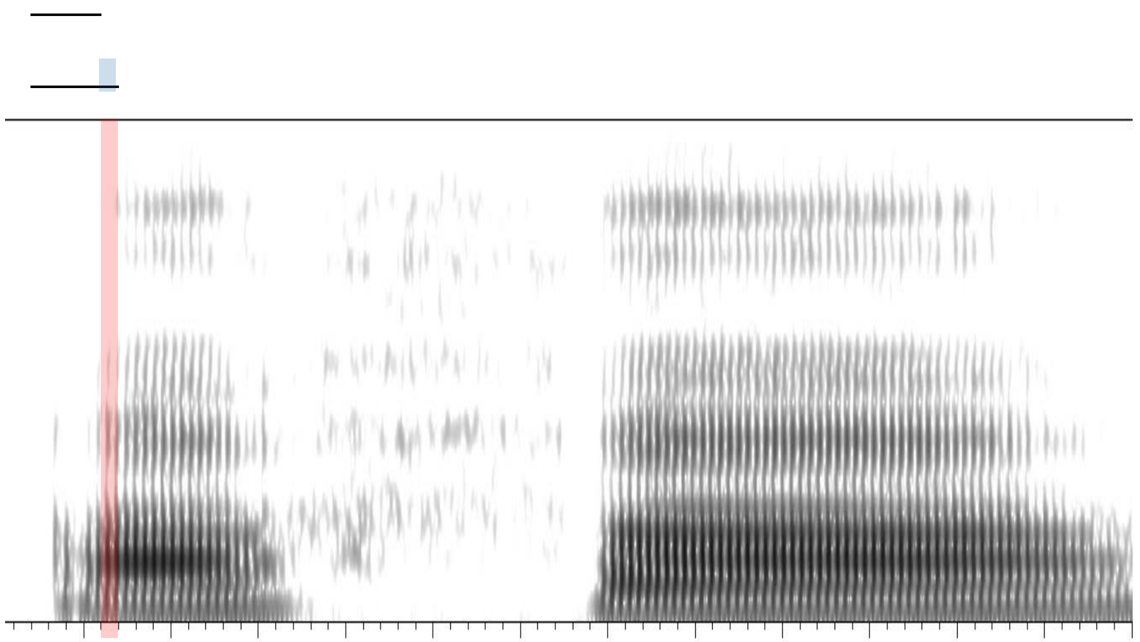




End-to-End Trained Automatic Speech Recognition

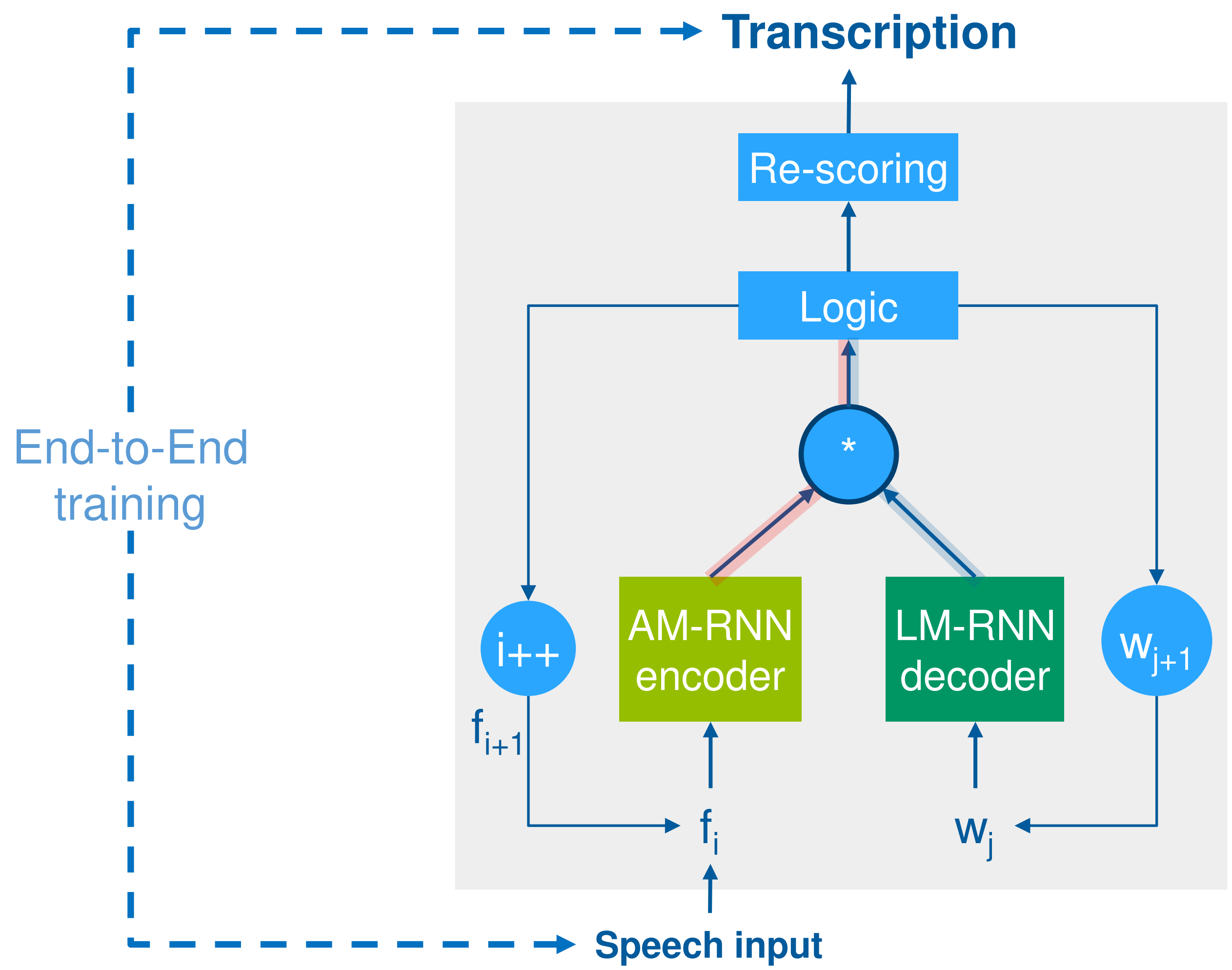
Recurrent Neural Network Transducer

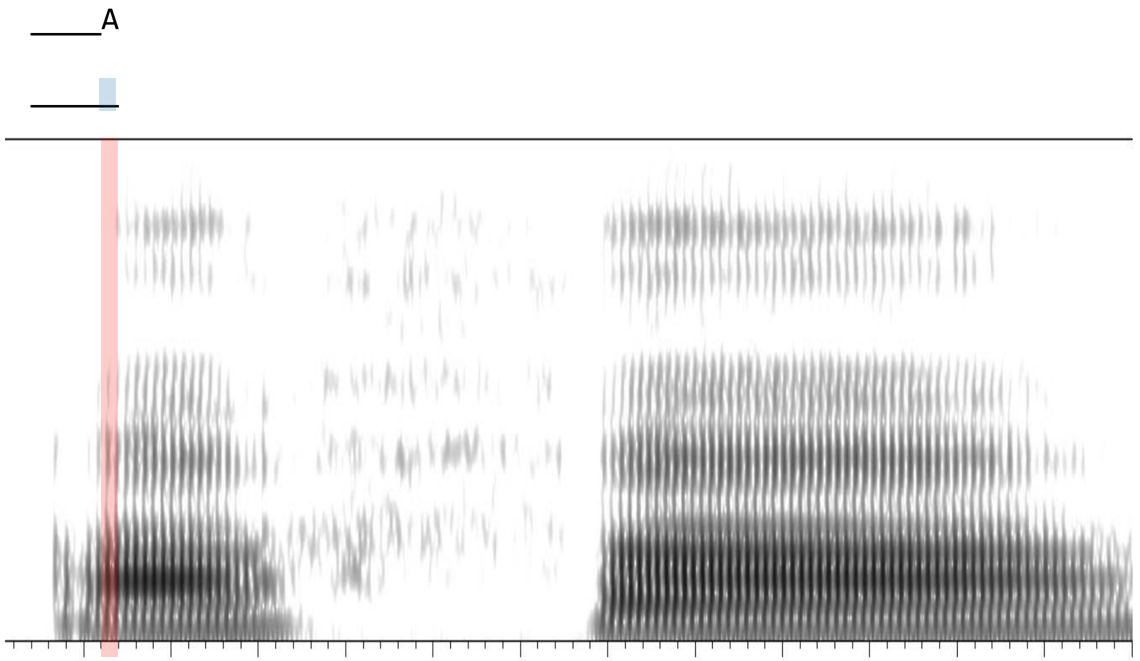




End-to-End Trained Automatic Speech Recognition

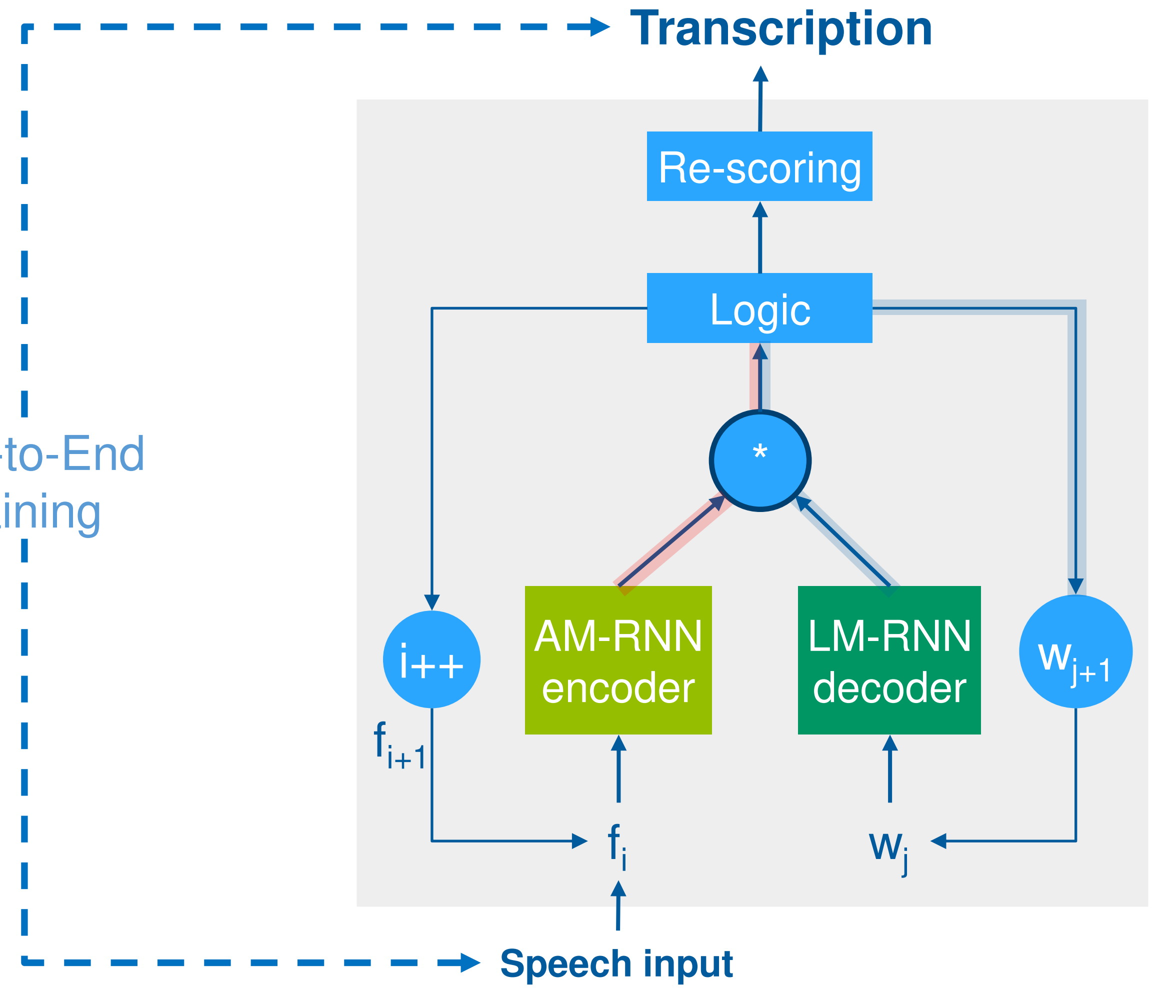
Recurrent Neural Network Transducer

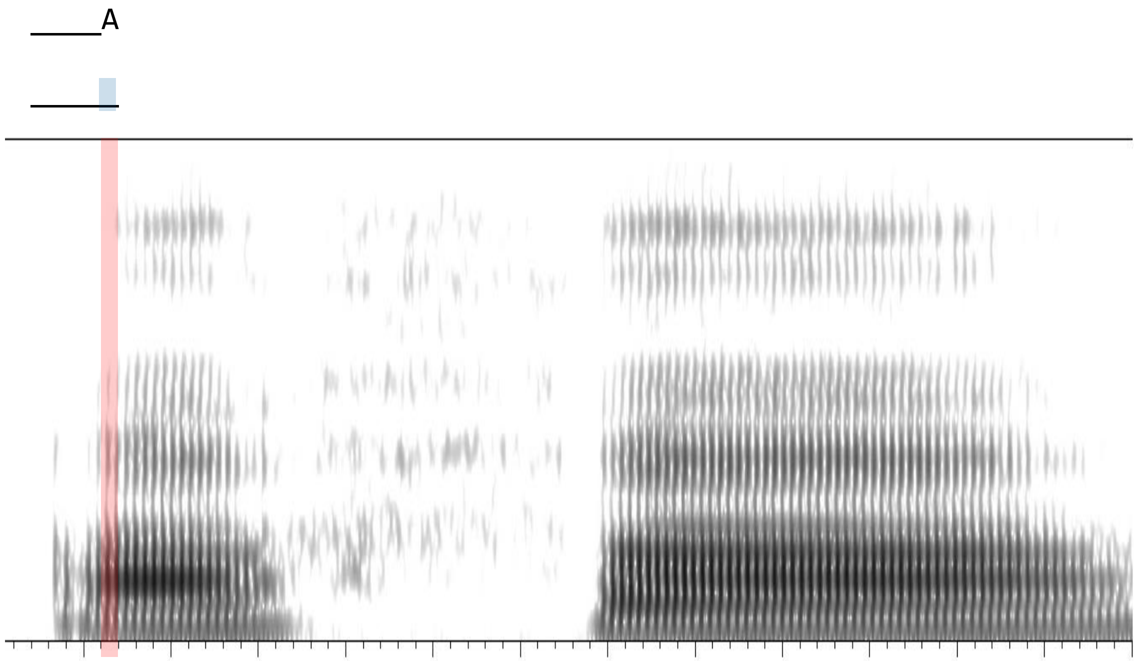




End-to-End Trained Automatic Speech Recognition

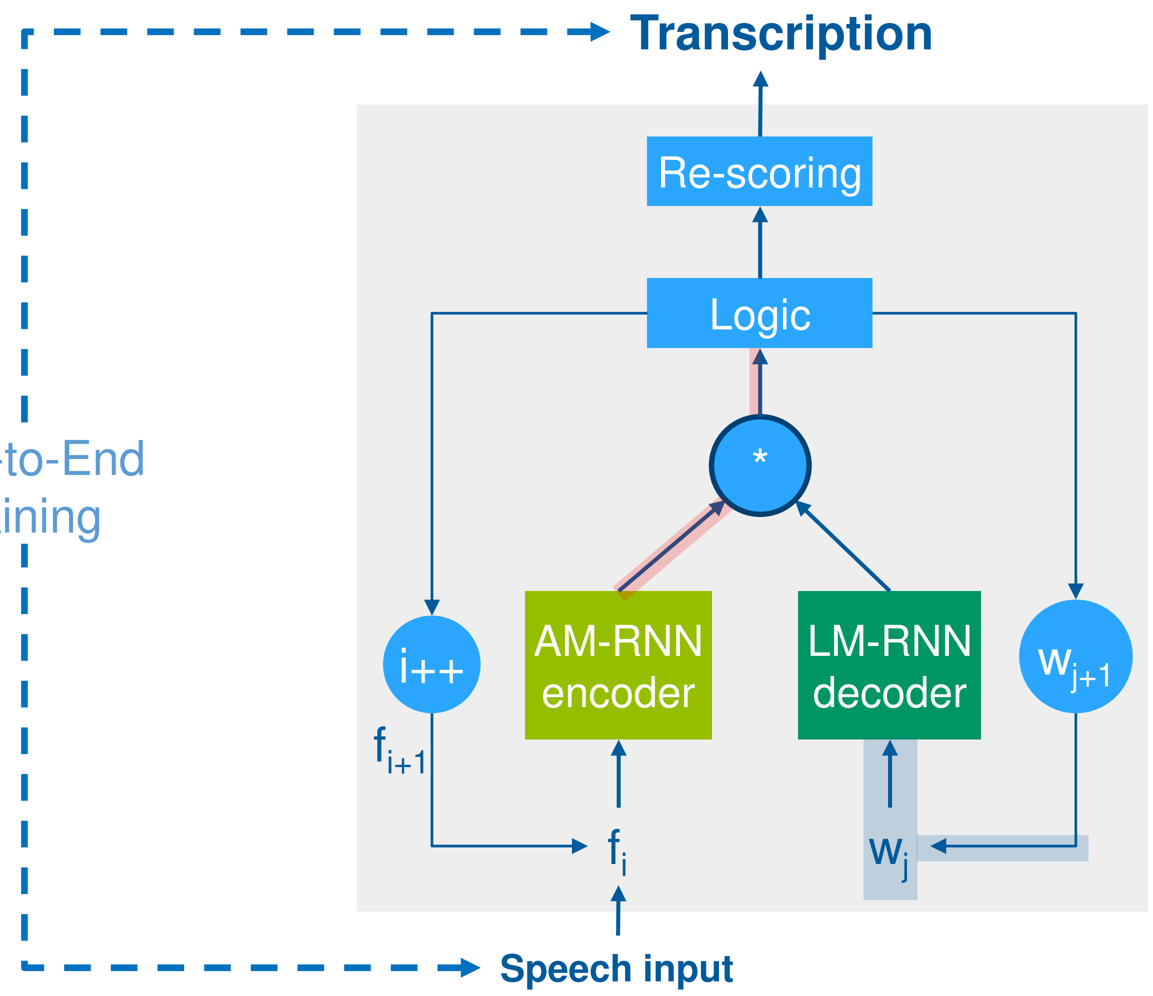
Recurrent Neural Network Transducer

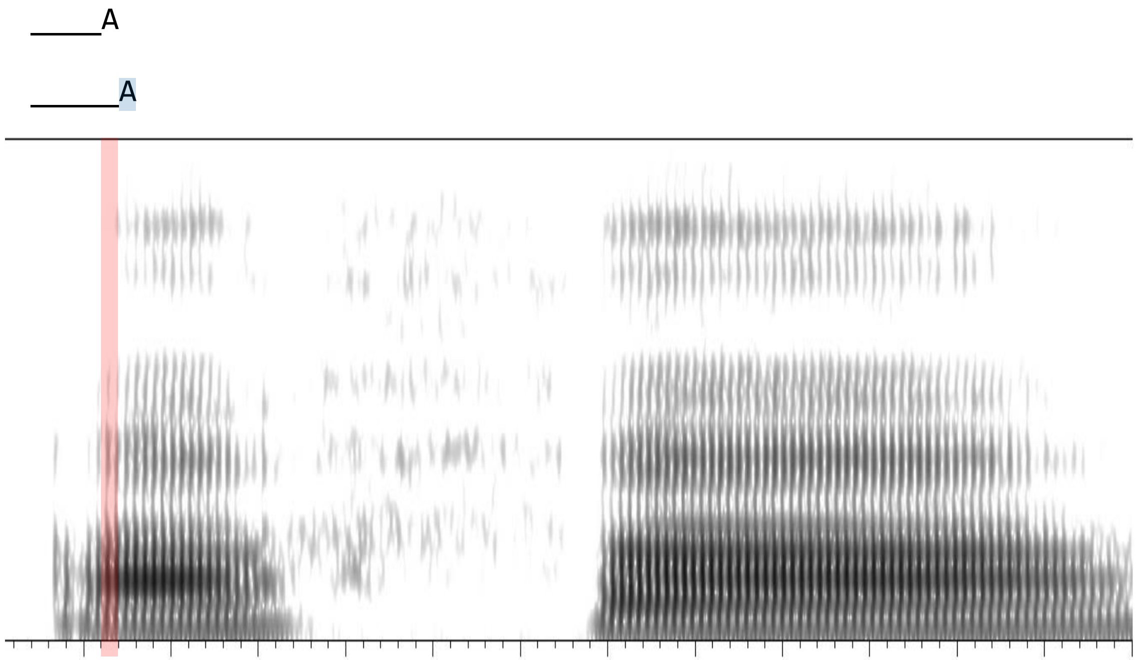




End-to-End Trained Automatic Speech Recognition

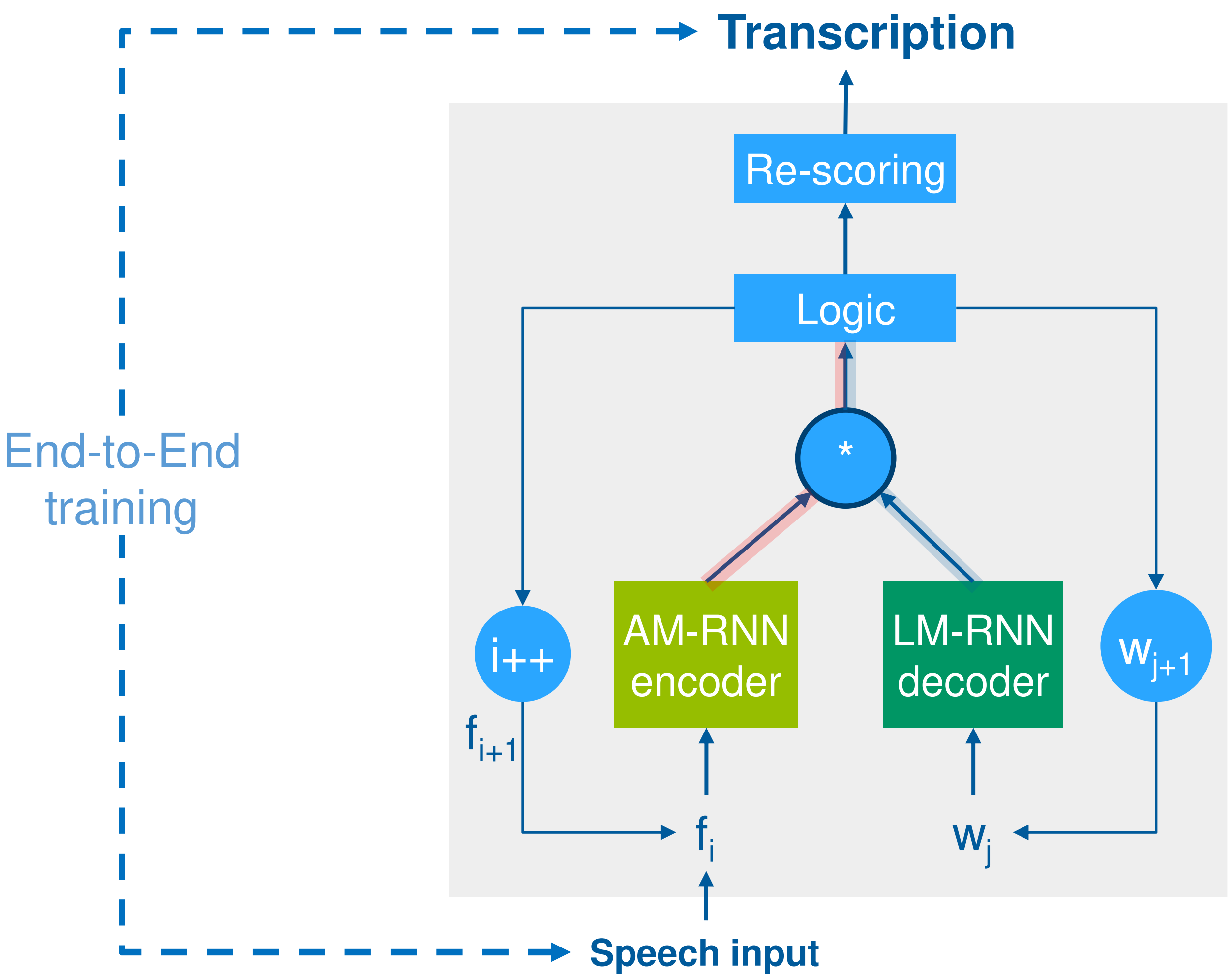
Recurrent Neural Network Transducer

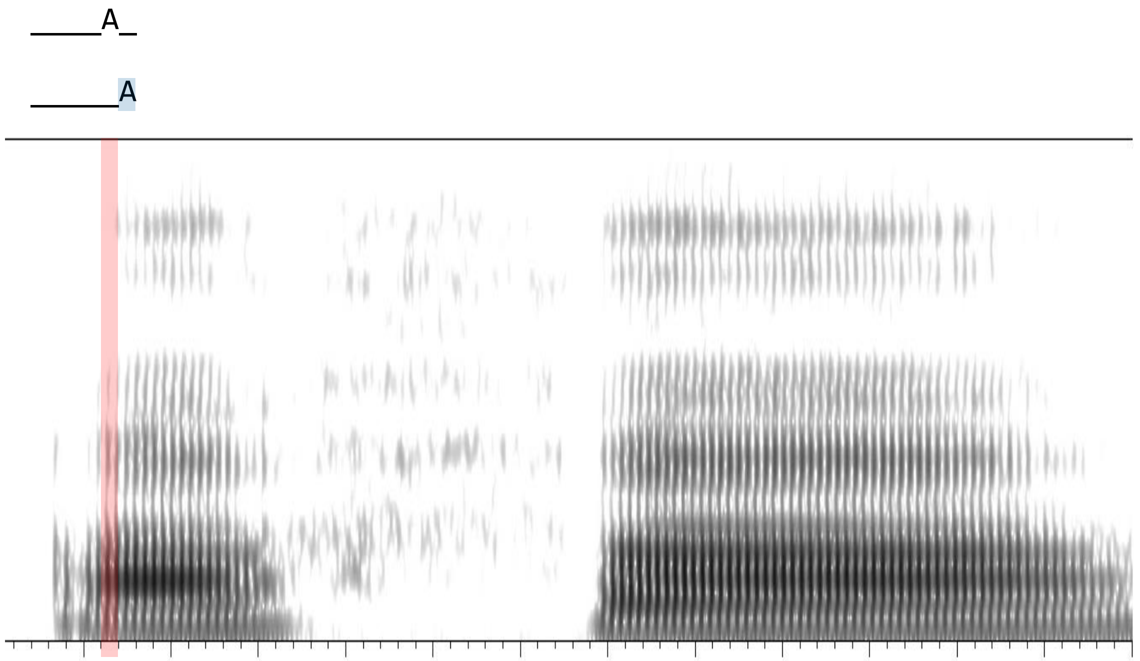




End-to-End Trained Automatic Speech Recognition

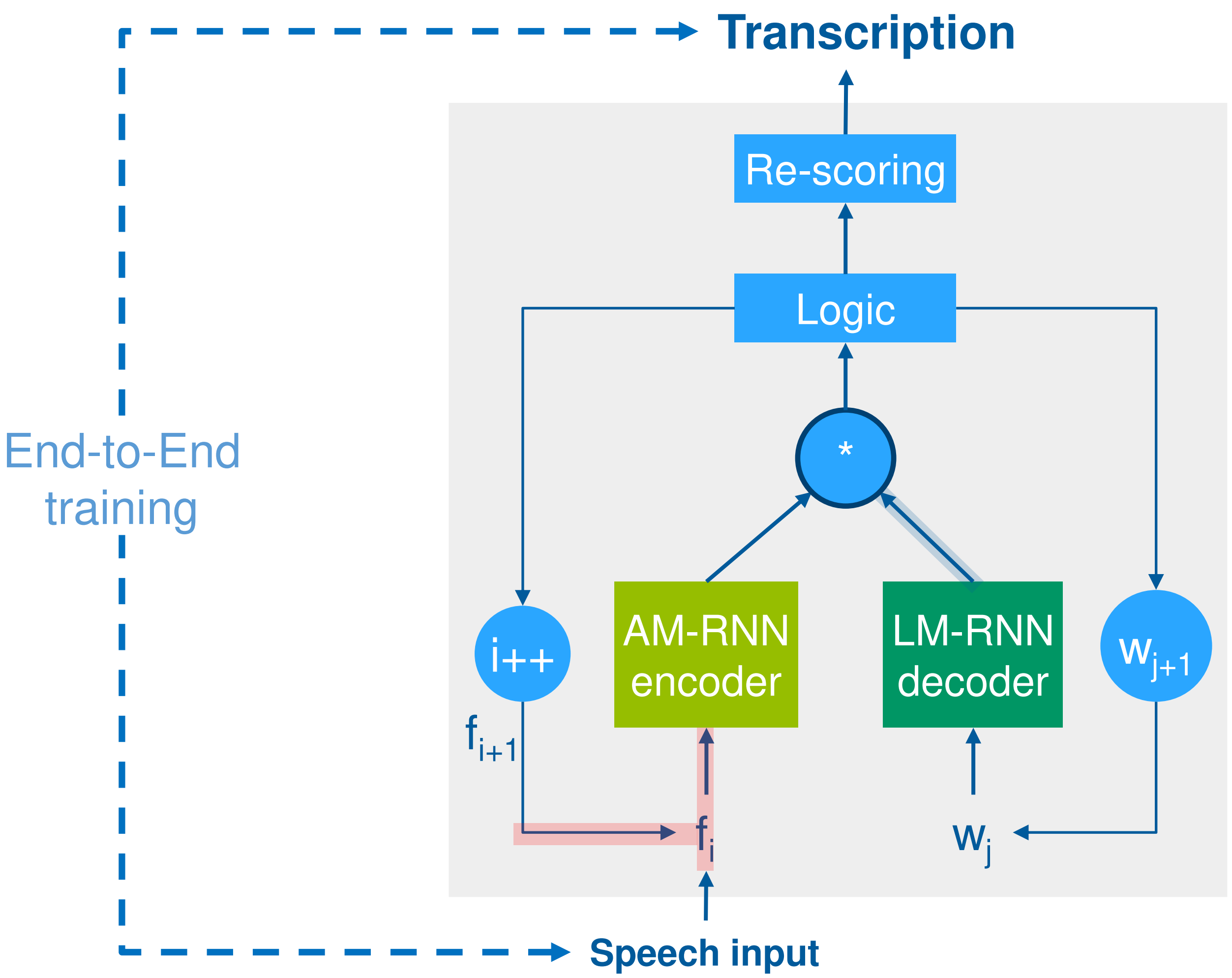
Recurrent Neural Network Transducer

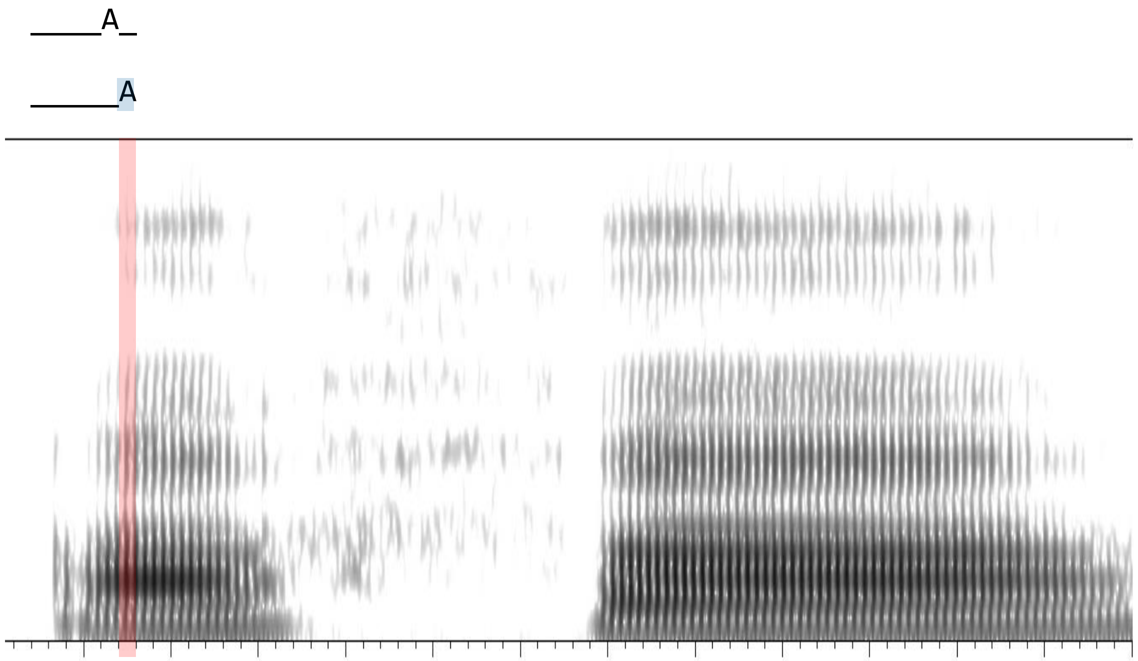




End-to-End Trained Automatic Speech Recognition

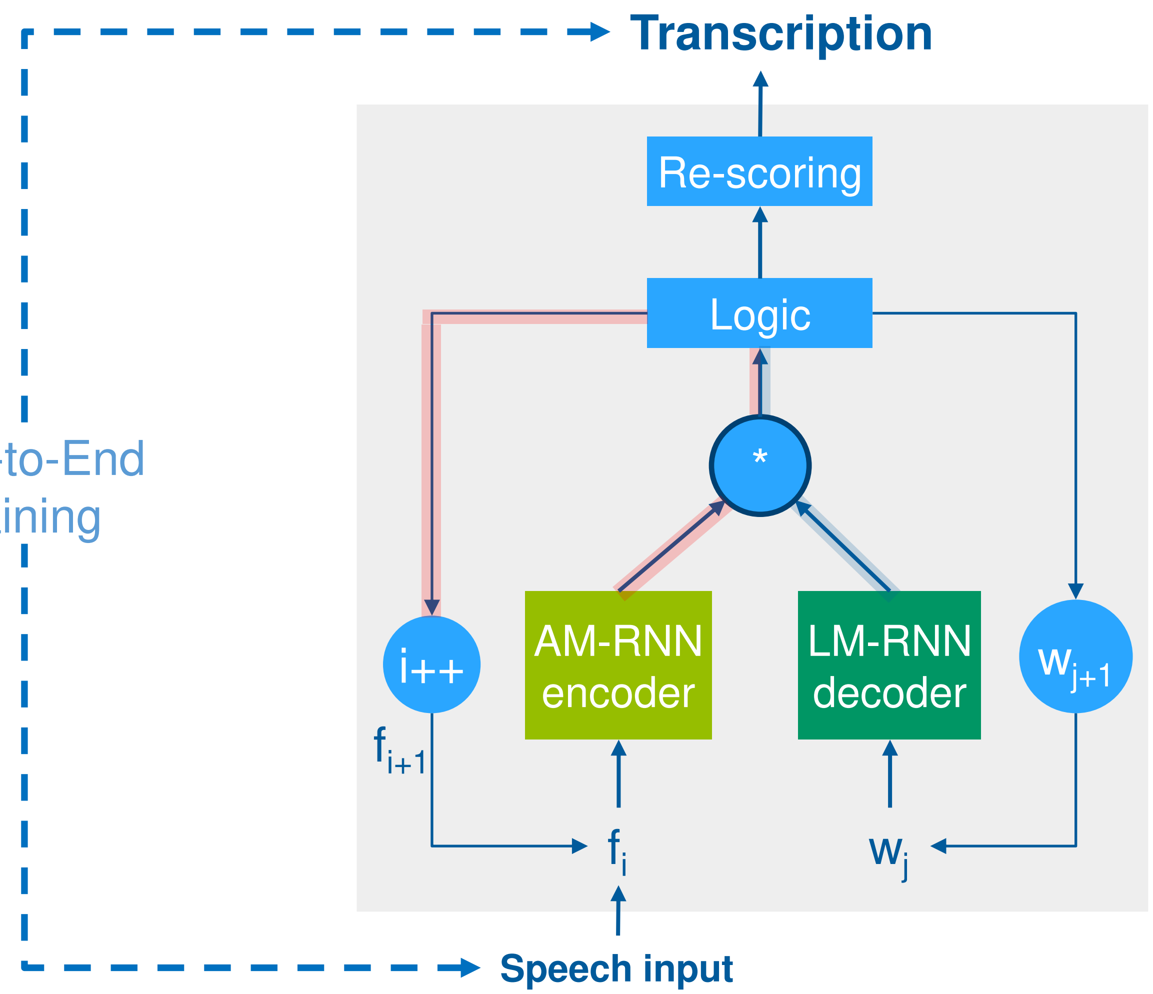
Recurrent Neural Network Transducer

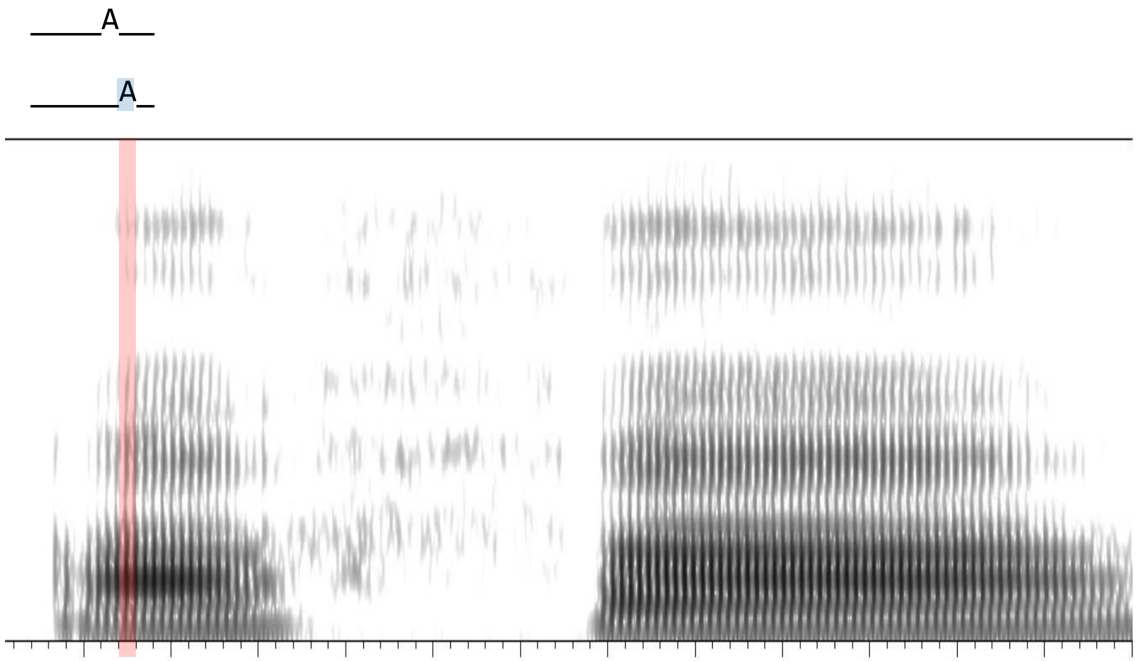




End-to-End Trained Automatic Speech Recognition

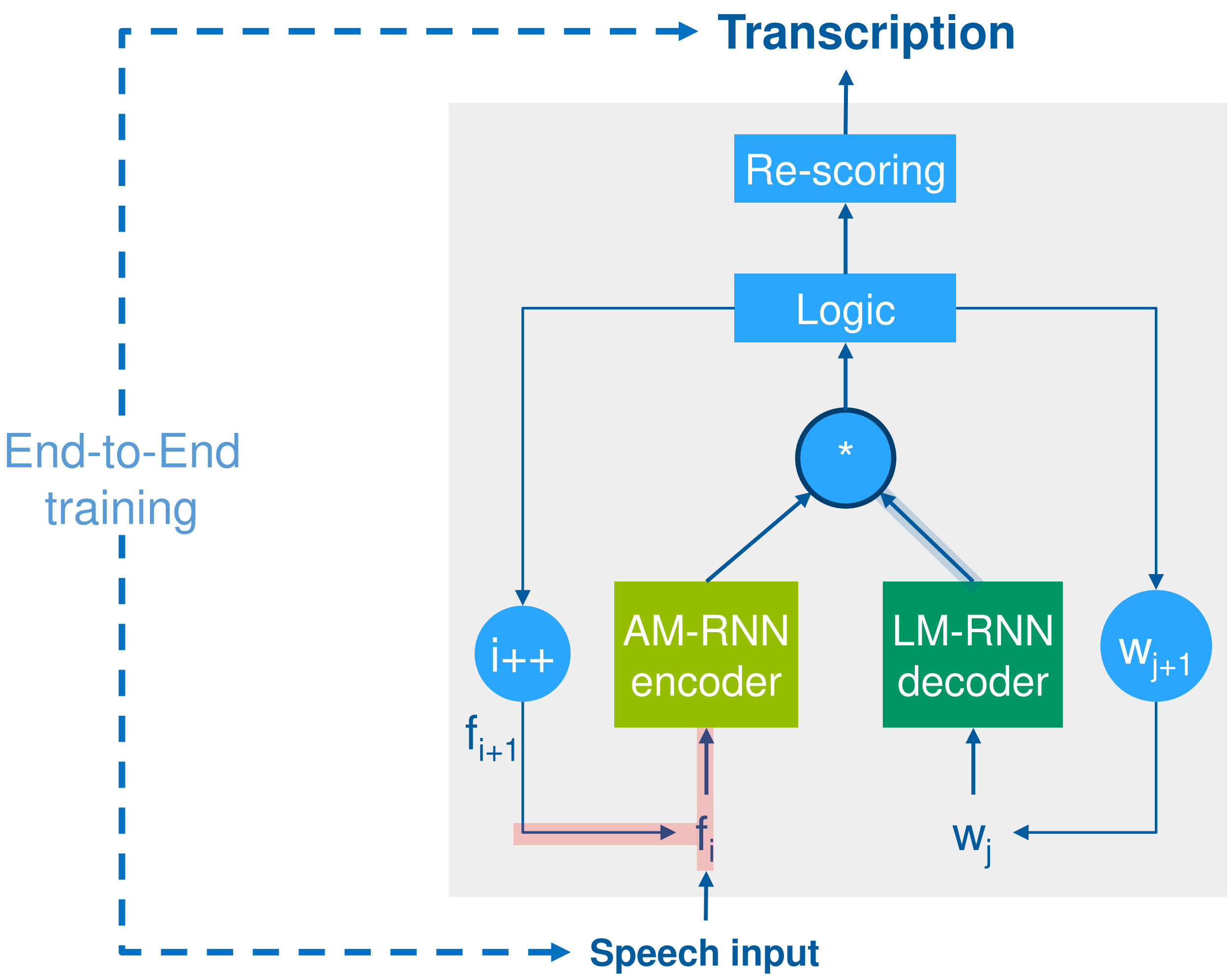
Recurrent Neural Network Transducer

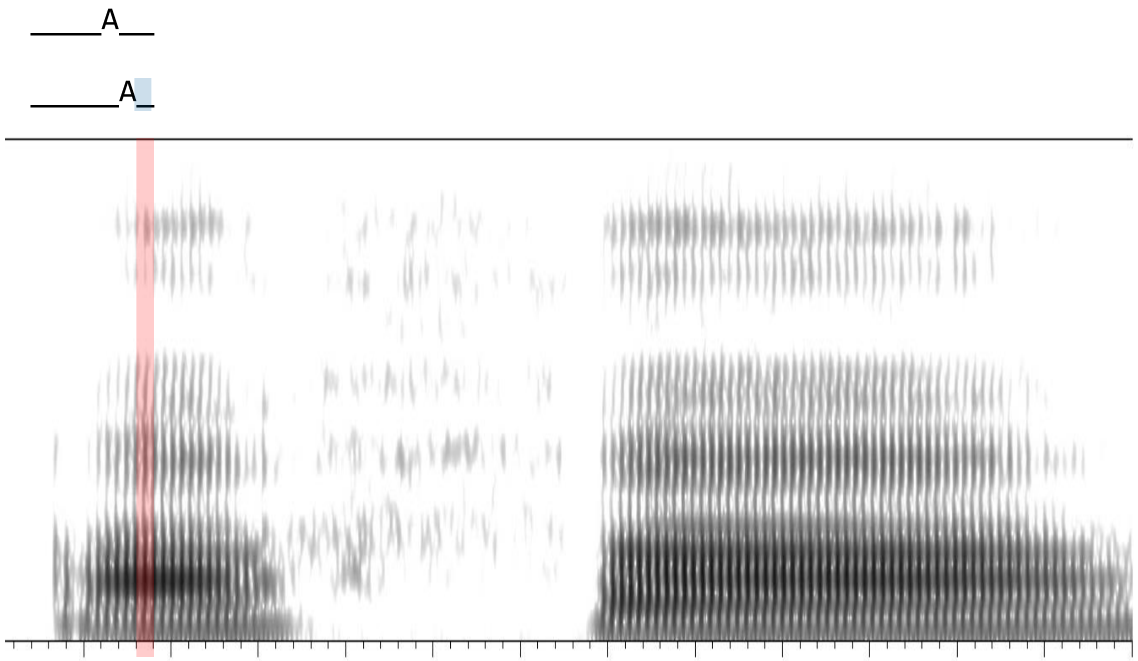




End-to-End Trained Automatic Speech Recognition

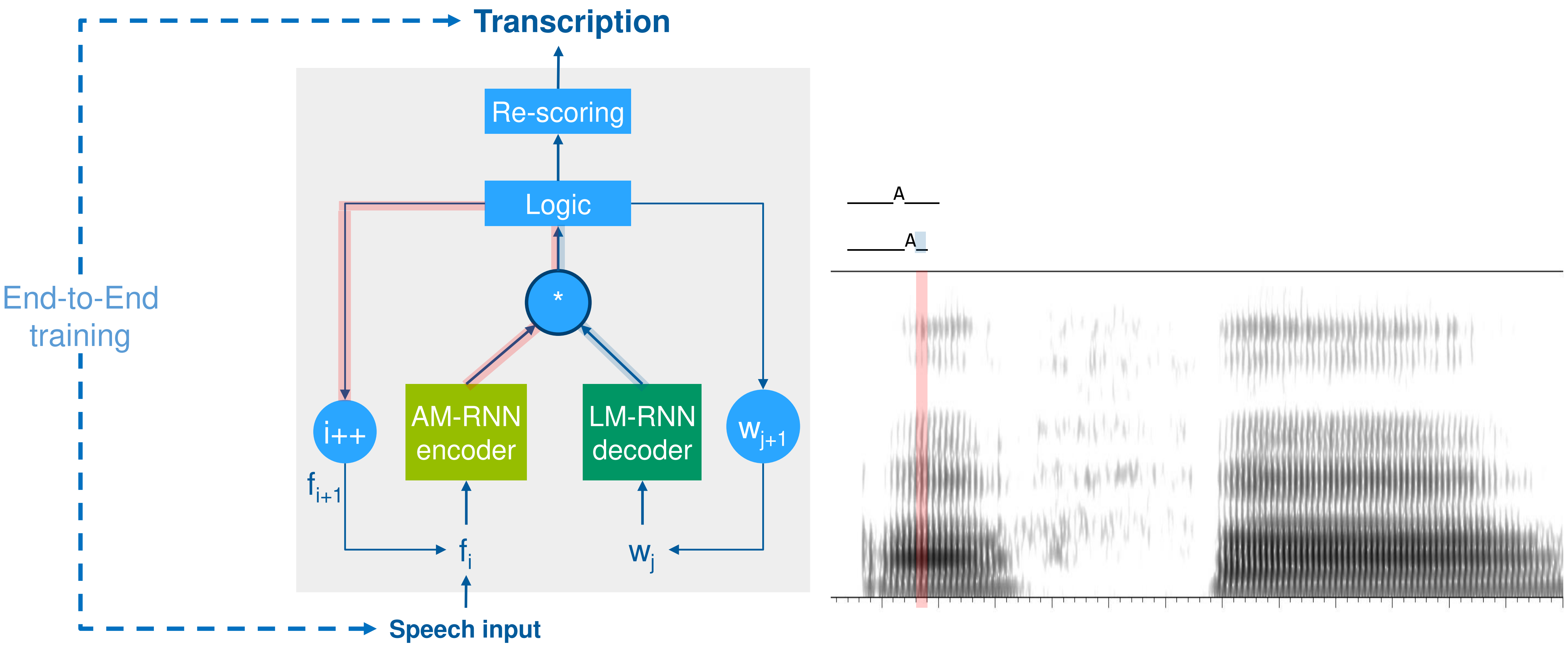
Recurrent Neural Network Transducer





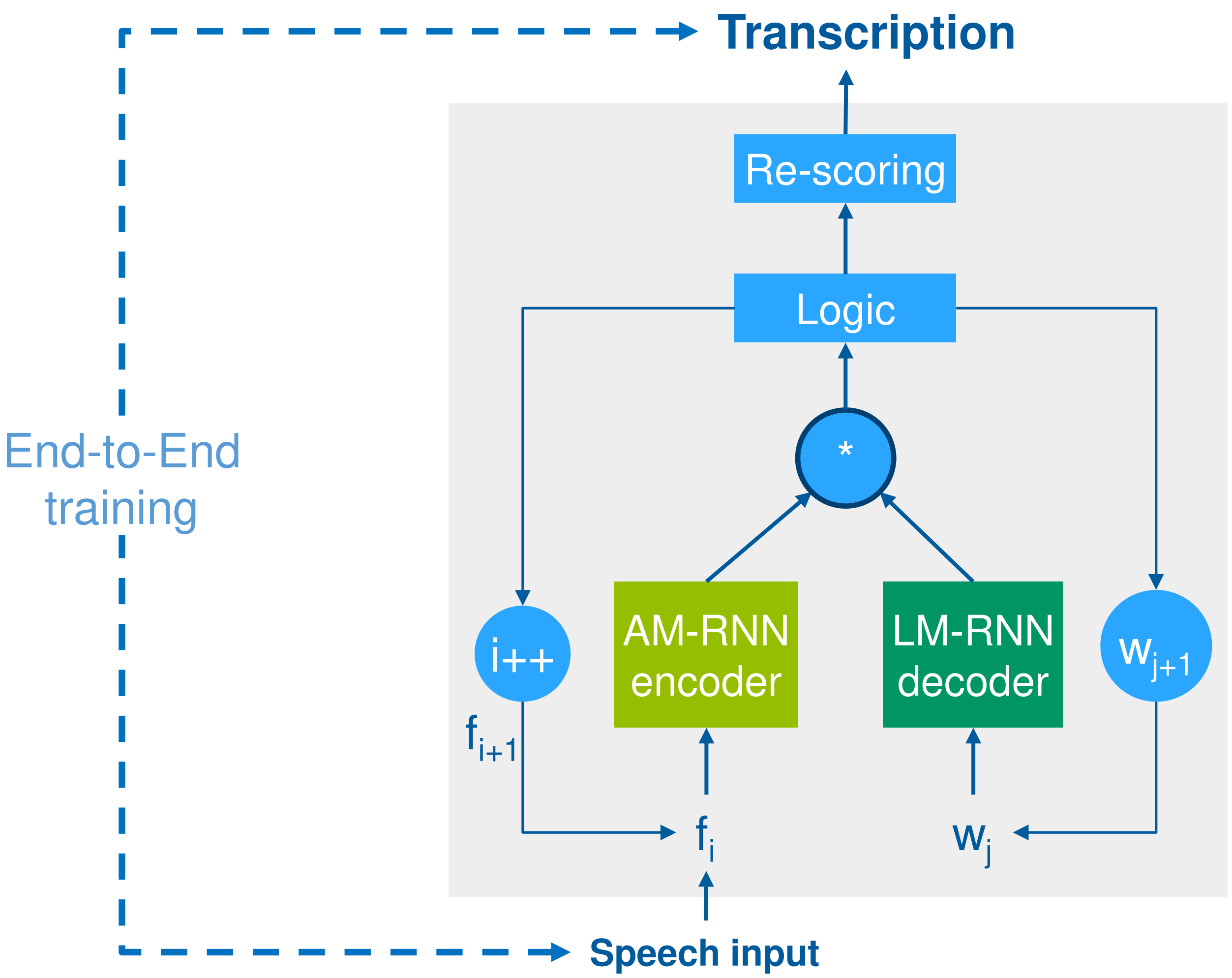
End-to-End Trained Automatic Speech Recognition

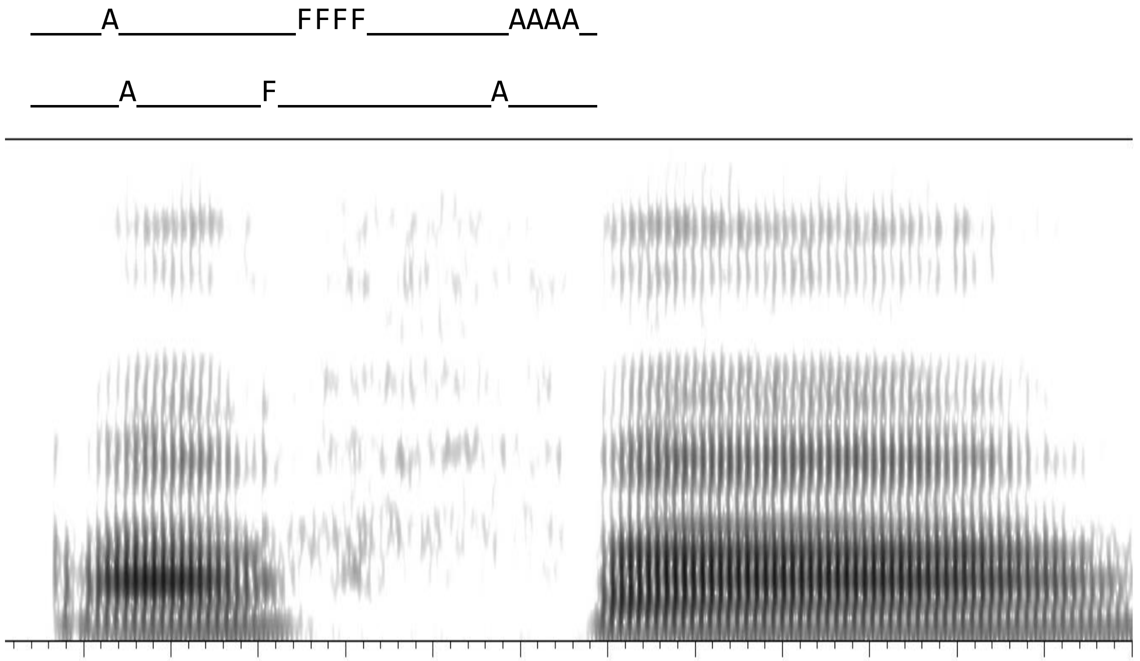
Recurrent Neural Network Transducer



End-to-End Trained Automatic Speech Recognition

Recurrent Neural Network Transducer





Spoken Language Understanding

Where we are, where we go.

Supports the user to find solutions

Mobile/laptop,
web search

Text/Voice query

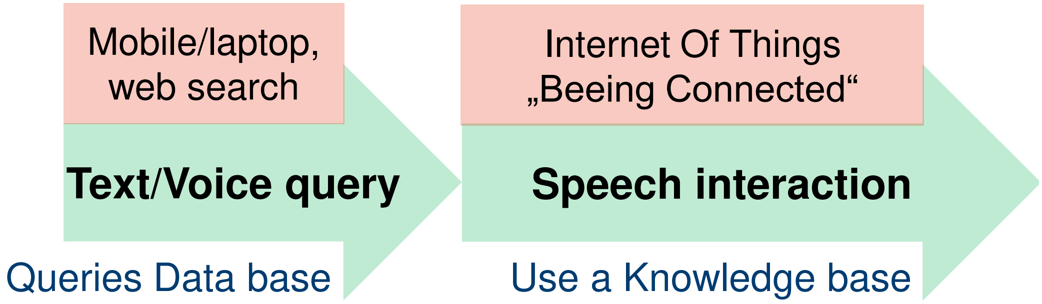
Queries Data base

Spoken Language Understanding

Where we are, where we go.

Supports the user to find solutions

Helps the user to solve his problems



Spoken Language Understanding

Where we are, where we go.

Supports the user to find solutions

Helps the user to solve his problems

Jointly find solution for
Problems the user has

Text/Voice query	Speech interaction	Personal Assistant
Mobile/laptop, web search	Internet Of Things „Beeing Connected“	Ambient Intelligence „Beeing Entertained“
Queries Data base	Use a Knowledge base	Makes conclusions

Spoken Language Understanding

Where we are, where we go.

Supports the user to find solutions

Helps the user to solve his problems

Jointly find solution for
Problems the user has

Find the solution for
Problems the user may have

Mobile/laptop, web search	Internet Of Things „Beeing Connected“	Ambient Intelligence „Beeing Entertained“	?? „Survice in a Complex World“
Text/Voice query	**Speech interaction**	**Personal Assistant**	**Communicate to solve problems**
Queries Data base	Use a Knowledge base	Makes conclusions	Has intuition

Spoken Language Understanding

Where we are, where we go.

Supports the user to find solutions

Helps the user to solve his problems

Jointly find solution for
Problems the user has

Find the solution for
Problems the user may have

Mobile/laptop, web search	Internet Of Things „Beeing Connected“	Ambient Intelligence „Beeing Entertained“	?? „Survice in a Complex World“
Text/Voice query	**Speech interaction**	**Personal Assistant**	**Communicate to solve problems**
Queries Data base	Use a Knowledge base	Makes conclusions	Has intuition

Search

Spoken Language Understanding

Where we are, where we go.

Supports the user to find solutions	Helps the user to solve his problems	Jointly find solution for Problems the user has	Find the solution for Problems the user may have
Mobile/laptop, web search	Internet Of Things „Beeing Connected“	Ambient Intelligence „Beeing Entertained“	?? „Survice in a Complex World“
Text/Voice query	**Speech interaction**	**Personal Assistant**	**Communicate to solve problems**
Queries Data base	Use a Knowledge base	Makes conclusions	Has intuition

Knowlege

Search

Spoken Language Understanding

Where we are, where we go.

Supports the user to find solutions

Helps the user to solve his problems

Jointly find solution for
Problems the user has

Find the solution for
Problems the user may have

Mobile/laptop,
web search

Text/Voice query

Queries Data base

Internet Of Things
„Beeing Connected“

Speech interaction

Use a Knowledge base

Ambient Intelligence
„Beeing Entertained“

Personal Assistant

Makes conclusions

??
„Survice in a Complex World“

Communicate to solve problems

Has intuition

Q&A

Knowlege

Search

Spoken Language Understanding

Where we are, where we go.

Supports the user to find solutions

Helps the user to solve his problems

Jointly find solution for
Problems the user has

Find the solution for
Problems the user may have

Mobile/laptop, web search	Internet Of Things „Beeing Connected“	Ambient Intelligence „Beeing Entertained“	?? „Survice in a Complex World“
Text/Voice query	**Speech interaction**	**Personal Assistant**	**Communicate to solve problems**
Queries Data base	Use a Knowledge base	Makes conclusions	Has intuition

Dialog

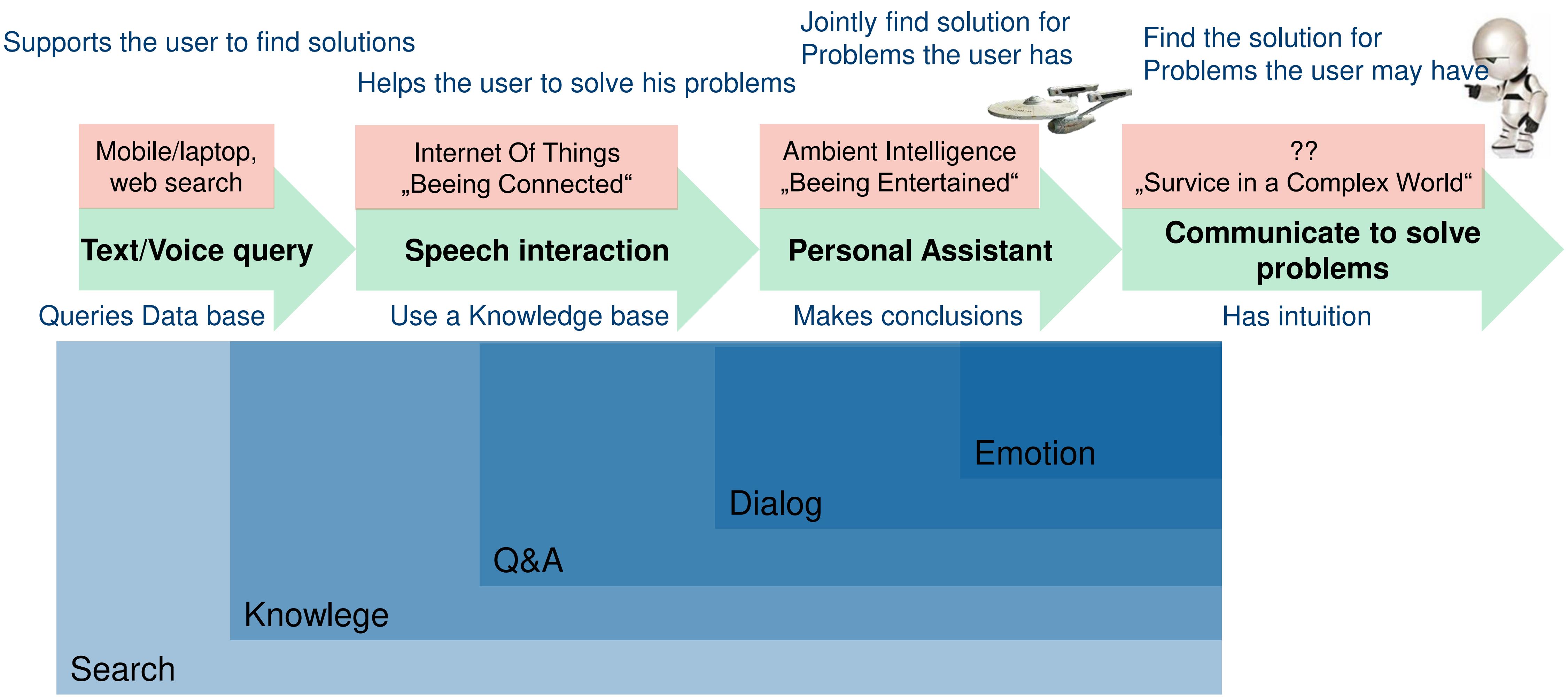
Q&A

Knowlege

Search

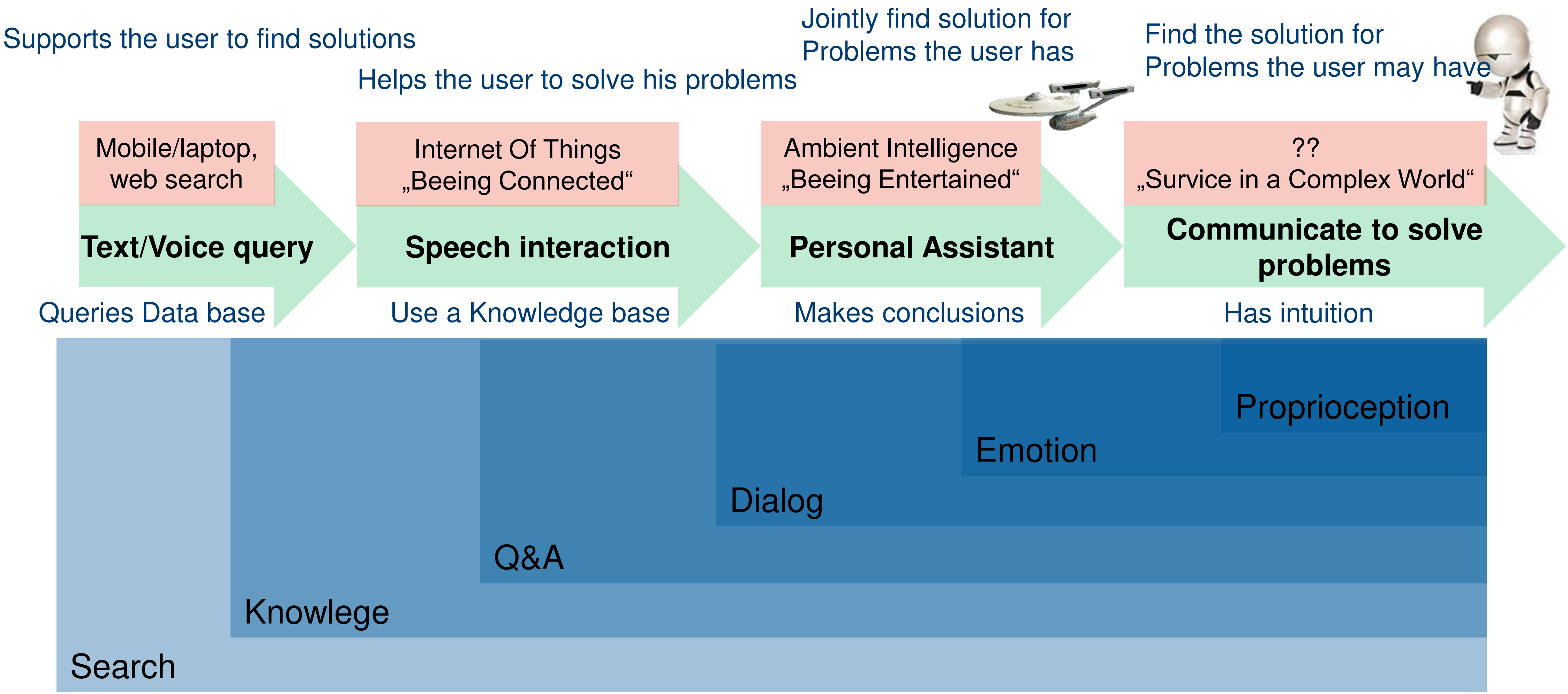
Spoken Language Understanding

Where we are, where we go.



Spoken Language Understanding

Where we are, where we go.



Spoken Language Understanding

Where we are, where we go.