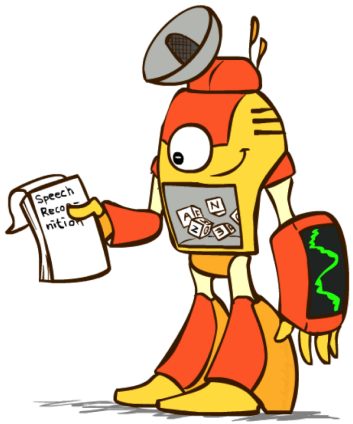
Introduction to Artificial Intelligence

Lecture 10: Communication

Prof. Gilles Louppe
g.louppe@uliege.be

LIÈGE université

Today



Can you talk to an artificial agent? Can it understand what you say?

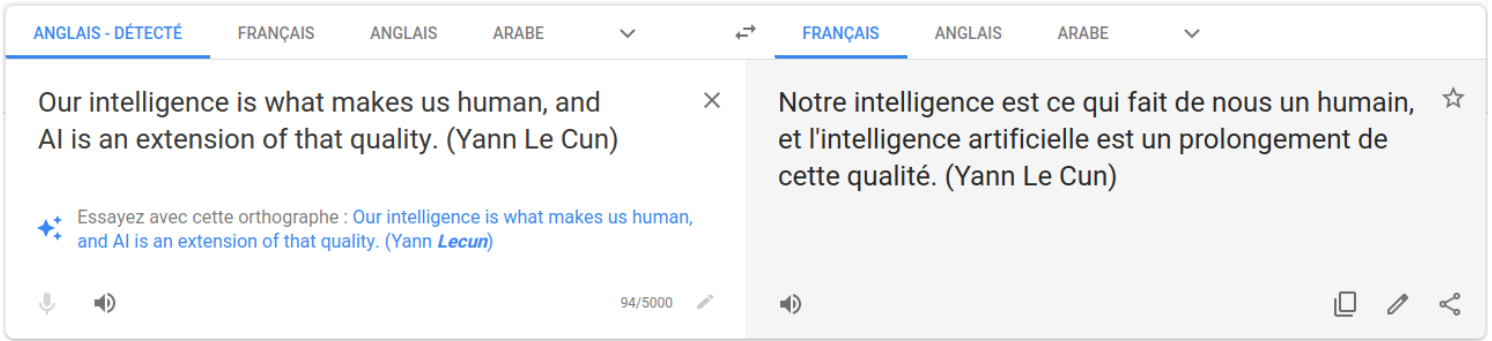
- Machine translation
- Speech recognition
- Text-to-speech synthesis

Image credits: CS188, UC Berkeley.

Sequence-to-sequence mapping

Machine translation:	Hello, my name is HAL.	→	Bonjour, mon nom est HAL.
Speech recognition:		→	Hello, my name is HAL.
Text-to-speech synthesis:	Hello, my name is HAL.	→	

Machine translation



Machine translation

Automatic translation of text from one natural language (the source) to another (the target), while preserving the intended meaning.

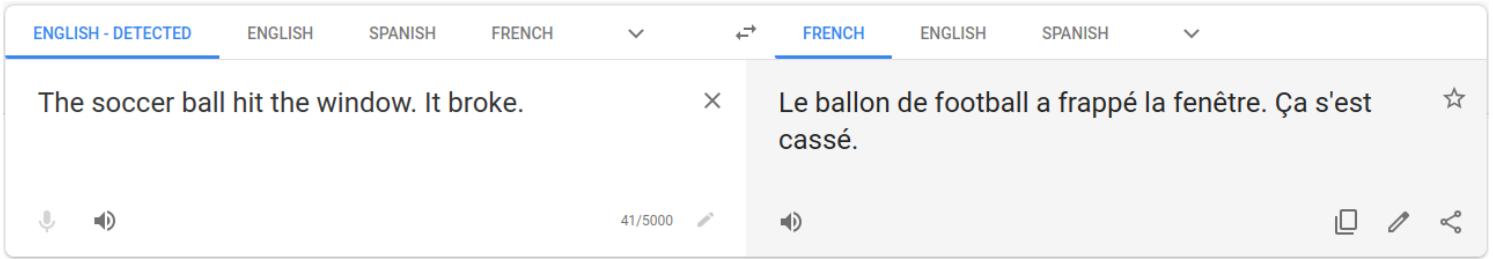
[Q] How would you engineer a machine translation system?

Issue of dictionary lookups

顶部	/**top**/roof/
顶端	/summit/peak/**top**/apex/
顶头	/coming directly towards one/**top**/end/
盖	/lid/**top**/cover/canopy/build/Gai/
盖帽	/surpass/**top**/
极	/extremely/pole/utmost/**top**/collect/receive/
尖峰	/peak/**top**/
面	/fade/side/surface/aspect/**top**/face/flour/
摘心	/**top**/topping/

Natural languages are not 1:1 mappings of each other!

Image credits: CS188, UC Berkeley.



To obtain a correct translation, one must decide whether "it" refers to the soccer ball or to the window.

Therefore, one must understand physics as well as language.

History

Warren Weaver

When I look at an article in Russian, I say: "This is really written in English, but it has been coded in some strange symbols. I will now proceed to decode."

John Pierce

"Machine Translation" presumably means going by algorithm from machine-readable source text to useful target text... In this context, there has been no machine translation...

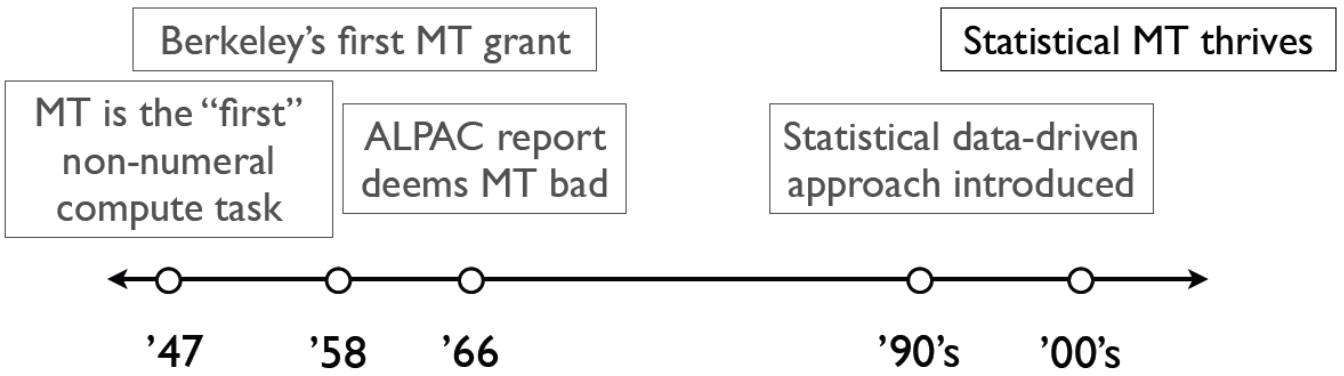


Image credits: CS188, UC Berkeley.

Data-driven machine translation

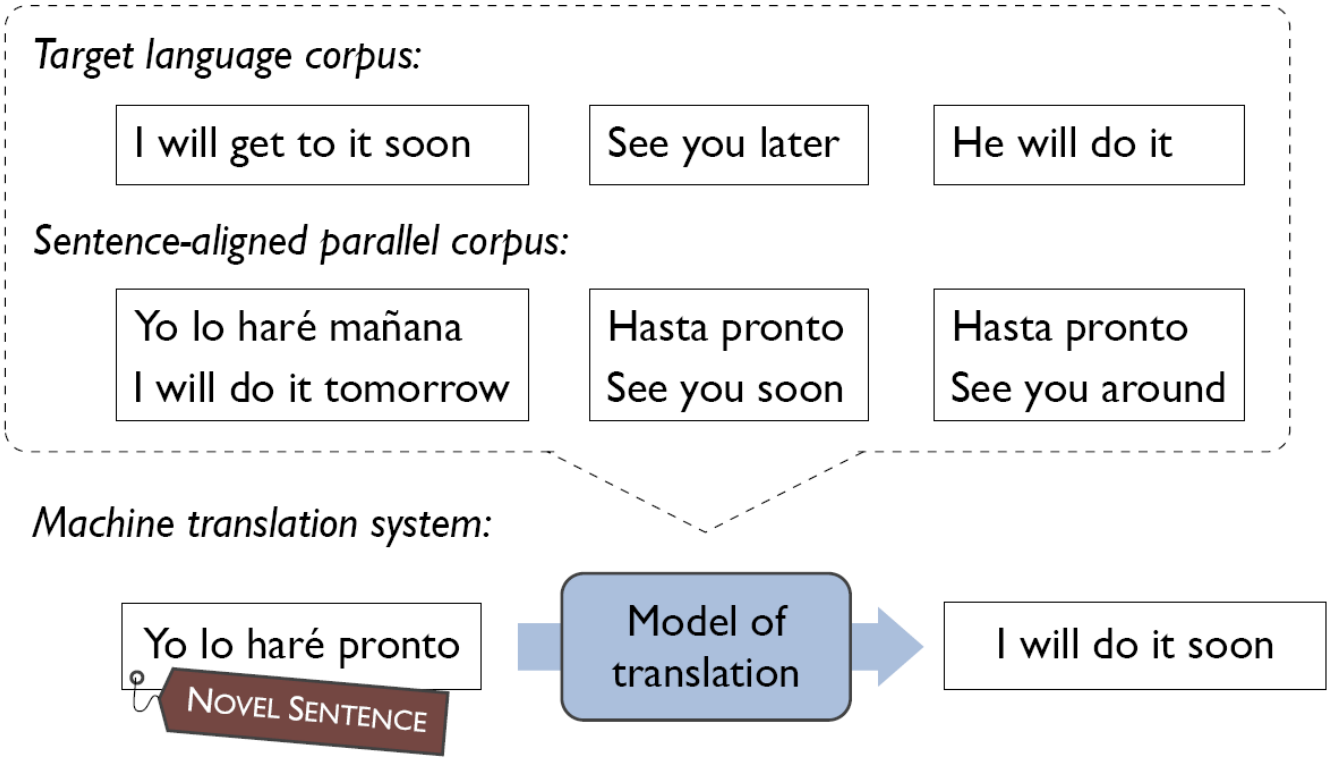


Image credits: CS188, UC Berkeley.

Machine translation systems

Translation systems must model the source and target languages, but systems vary in the type of models they use.

- Some systems analyze the source language text all the way into an interlingua knowledge representation and then generate sentences in the target language from that representation.
- Other systems are based on a transfer model. They keep a database of translation rules and whenever the rule matches, they translate directly. Transfer can occur at the lexical, syntactic or semantic level.

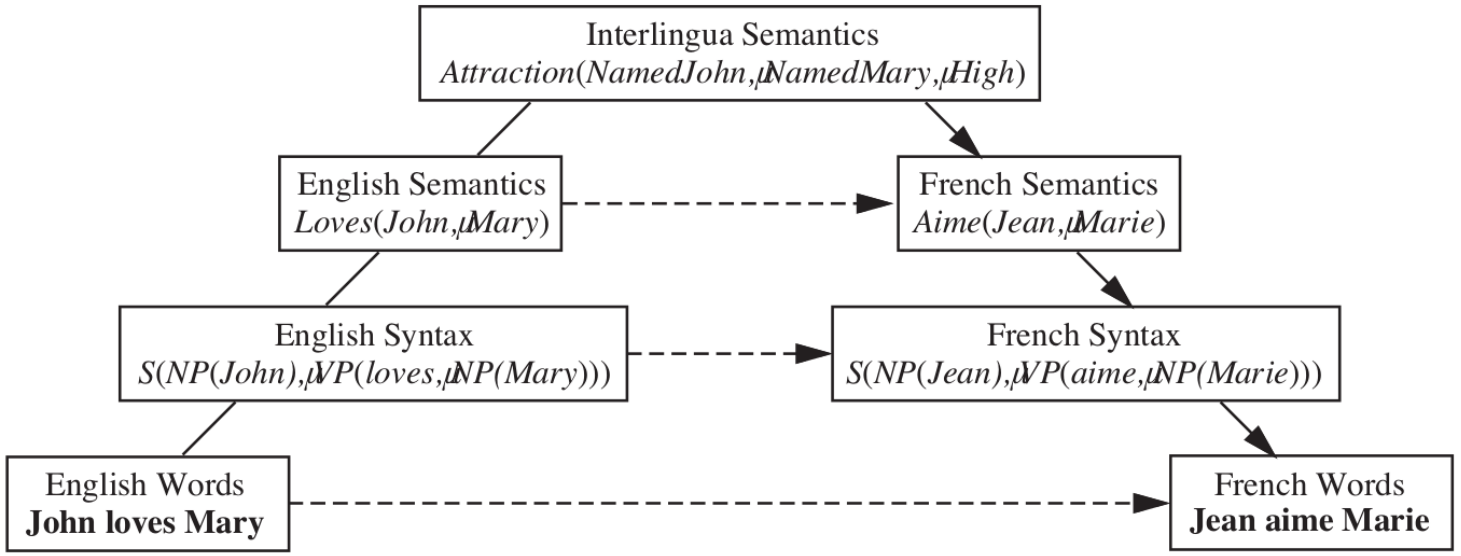


Figure 23.12 The Vauquois triangle: schematic diagram of the choices for a machine translation system (Vauquois, 1968). We start with English text at the top. An interlingua-based system follows the solid lines, parsing English first into a syntactic form, then into a semantic representation and an interlingua representation, and then through generation to a semantic, syntactic, and lexical form in French. A transfer-based system uses the dashed lines as a shortcut. Different systems make the transfer at different points; some make it at multiple points.

Statistical machine translation

To translate an English sentence e into a French sentence f, we seek the strings of words f^* such that

$$f^* = \arg\max_f P(f|e).$$

- The language model $P(f|e)$ is learned from a bilingual corpus, i.e. a collection of parallel texts, each an English/French pair.
- Most of the English sentences to be translated will be novel, but will be composed of phrases that that have been seen before.
- The corresponding French phrases will be reassembled to form a French sentence that makes sense.

Given an English source sentence e, finding a French translation f is a matter of three steps:

- Break e into phrases $e_1, \ldots, e_n$.
- For each phrase e_i, choose a corresponding French phrase f_i. We use the notation $P(f_i|e_i)$ for the phrasal probability that f_i is a translation of e_i.
- Choose a permutation of the phrases $f_1, \ldots, f_n$. For each f_i, we choose a distortion

$$d_i = \text{start}(f_i) - \text{end}(f_{i-1}) - 1,$$

  which is the number of words that phrase f_i has moved with respect to f_{i-1}; positive for moving to the right, negative for moving the left.

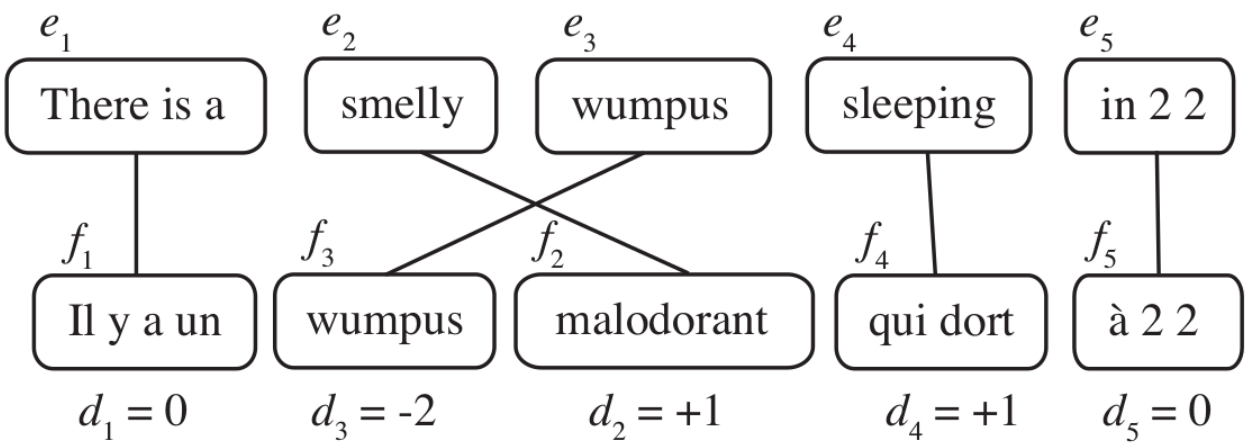


Figure 23.13 Candidate French phrases for each phrase of an English sentence, with distortion (d) values for each French phrase.

We define the probability $P(f, d|e)$ that the sequence of phrases f with distortions d is a translation of the sequence of phrases e.

Assuming that each phrase translation and each distortion is independent of the others, we have

$$P(f, d|e) = \prod_i P(f_i|e_i)P(d_i).$$

- The best f and e cannot be found through enumeration because of the combinatorial explosion.
- Instead, local beam search with a heuristic that estimates probability has proven effective at finding a nearly-most-probable translation.

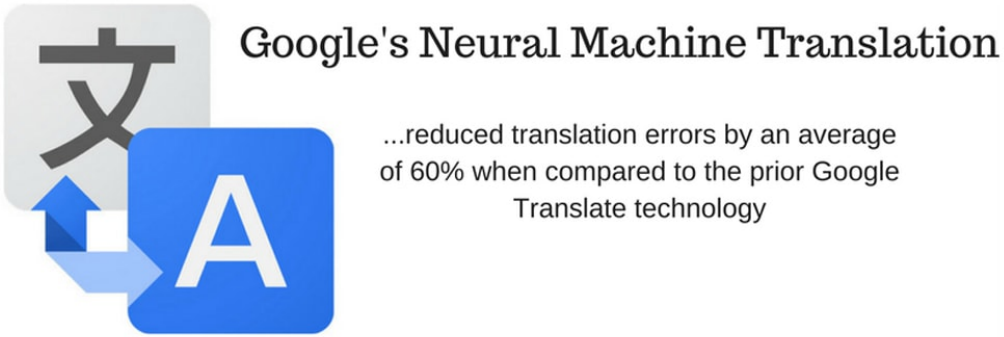
All that remains is to learn the phrasal and distortion probabilities:

1. Find parallel texts.
2. Segment into sentences.
3. Align sentences.
4. Align phrases.
5. Extract distortions.
6. Improve estimates with expectation-maximization.

Neural machine translation

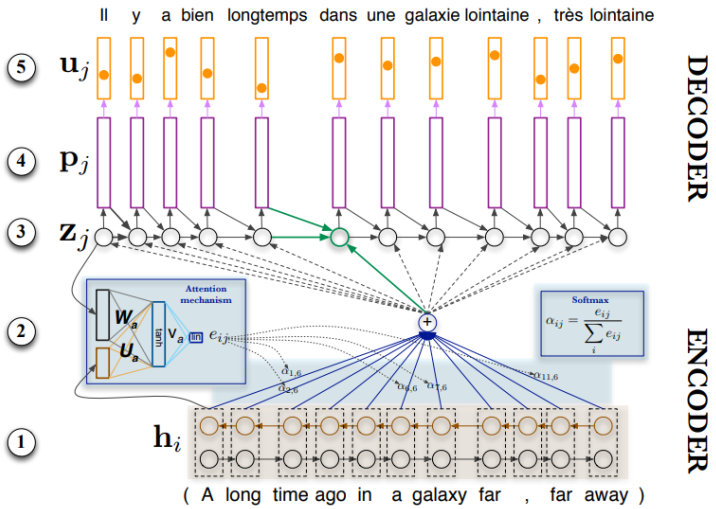
Modern machine translation systems are all based on neural networks of various types, often architectured as compositions of

- recurrent networks for sequence-to-sequence learning,
- convolutional networks for modeling spatial dependencies.



Attention-based recurrent neural network

- Encoder: bidirectional RNN, producing a set of annotation vectors h_i.
- Decoder: attention-based.
  - Compute attention weights α_{ij}.
  - Compute the weighted sum of the annotation vectors, as a way to align the input words to the output words.
  - Decode using the context vector, the embedding of the previous output word and the hidden state.



Unsupervised machine translation

- The latest approaches do not even need to have a bilingual corpus!
- Machine translation can be learned in a fully unsupervised way with unsupervised alignment.

Word-by-word translation:

- Learn a neural word embedding trained to predict the words around a given word using a context.
- Embedding in different languages share similar neighborhood structure.
- The system learn rotation of the word embedding in one language to match the word embedding in the other language, using adversarial training.
- This can be used to infer a fairly accurate bilingual dictionary without access to any translation!

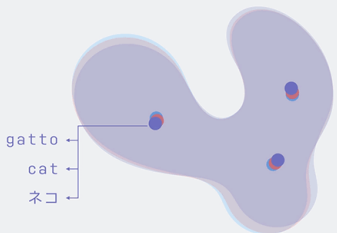


Image credits: Facebook AI Research, 2018. Unsupervised machine translation.

Translating sentences:

- Bootstrap the translation model with word-by-word initialization.
- The neural translation model must be able to reconstruct a sentence in a given language from a noisy version of it.
- The model also learns to reconstruct any source sentence given a noisy translation of the same sentence in the target domain, and vice-versa.
- The source and target sentence latent representations are constrained to have the same latent distributions through adversarial training.

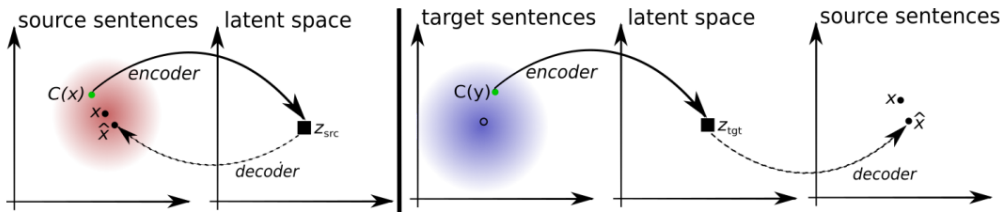


Figure 1: Toy illustration of the principles guiding the design of our objective function. Left (auto-encoding): the model is trained to reconstruct a sentence from a noisy version of it. x is the target, $C(x)$ is the noisy input, $\hat{x}$ is the reconstruction. Right (translation): the model is trained to translate a sentence in the other domain. The input is a noisy translation (in this case, from source-to-target) produced by the model itself, M, at the previous iteration (t), $y = M^{(t)}(x)$. The model is symmetric, and we repeat the same process in the other language. See text for more details.

Image credits: Lample et al, 2017. arXiv:1711.00043.

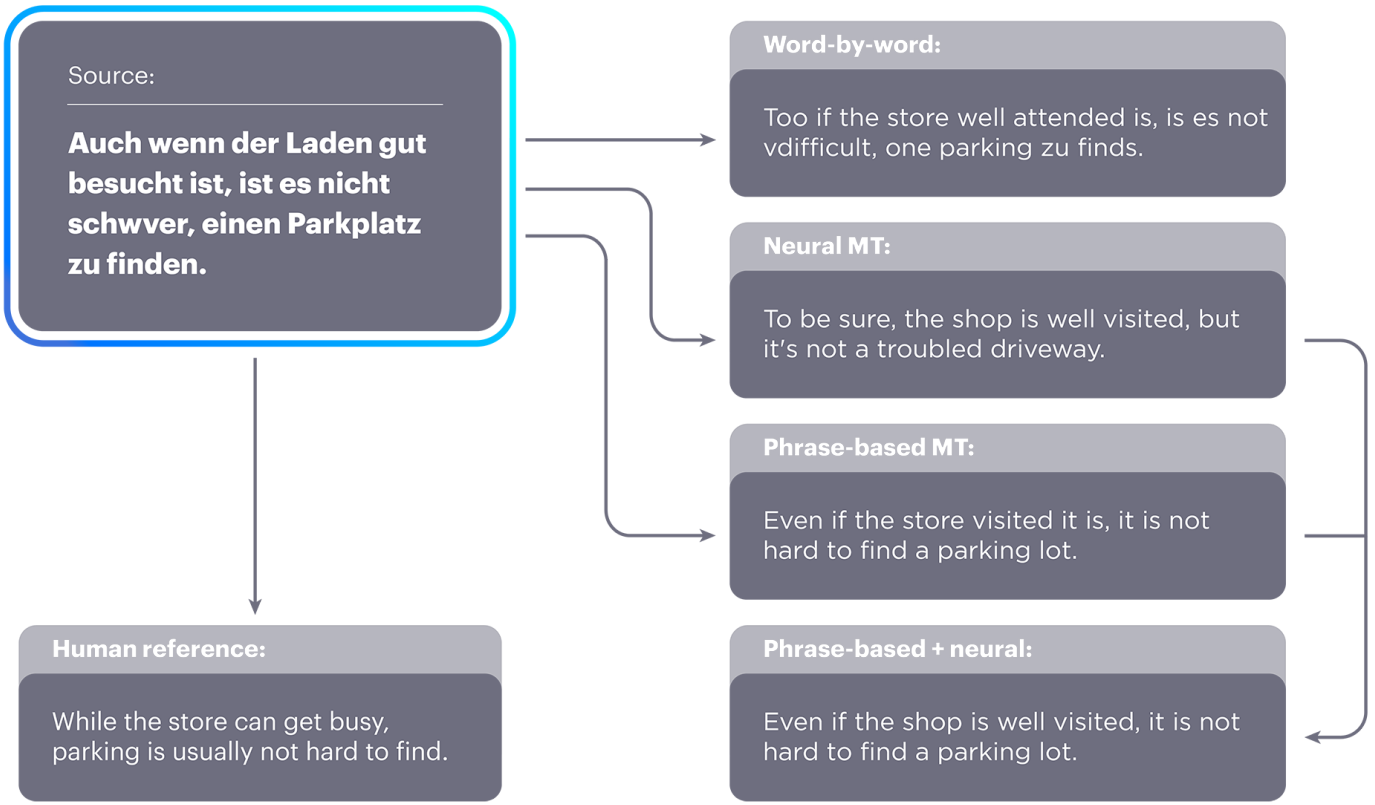


Image credits: Facebook AI Research, 2018. Unsupervised machine translation.

Speech recognition


Hey Google,
12:26 PM · Mon, July 30
72° Sunny
MON
76° 50°
TUE
67° 46°
WED
69° 48°
Coming Up
James · Event · July 30
7pm Salsa dancing class with Kate
Samantha · Reminder · Aug 11
9:30 am Soccer practice

Recognition as inference

$\mathbf{y}_{1:T}$

$\rightarrow$

$\mathbf{w}_{1:L}$

My name is HAL.

Speech recognition can be viewed as an instance of the problem of finding the most likely sequence of state variables $\mathbf{w}_{1:L}$, given a sequence of observations $\mathbf{y}_{1:T}$.

- In this case, (hidden) state variables are the words and the observations are sounds.
- The input audio waveform from a microphone is converted into a sequence of fixed size acoustic vectors $\mathbf{y}_{1:T}$ in a process called feature extraction.
- The decoder attempts to find the sequence of words $\mathbf{w}_{1:L} = w_1, \ldots, w_L$ which is the most likely to have generated $\mathbf{y}_{1:T}$:

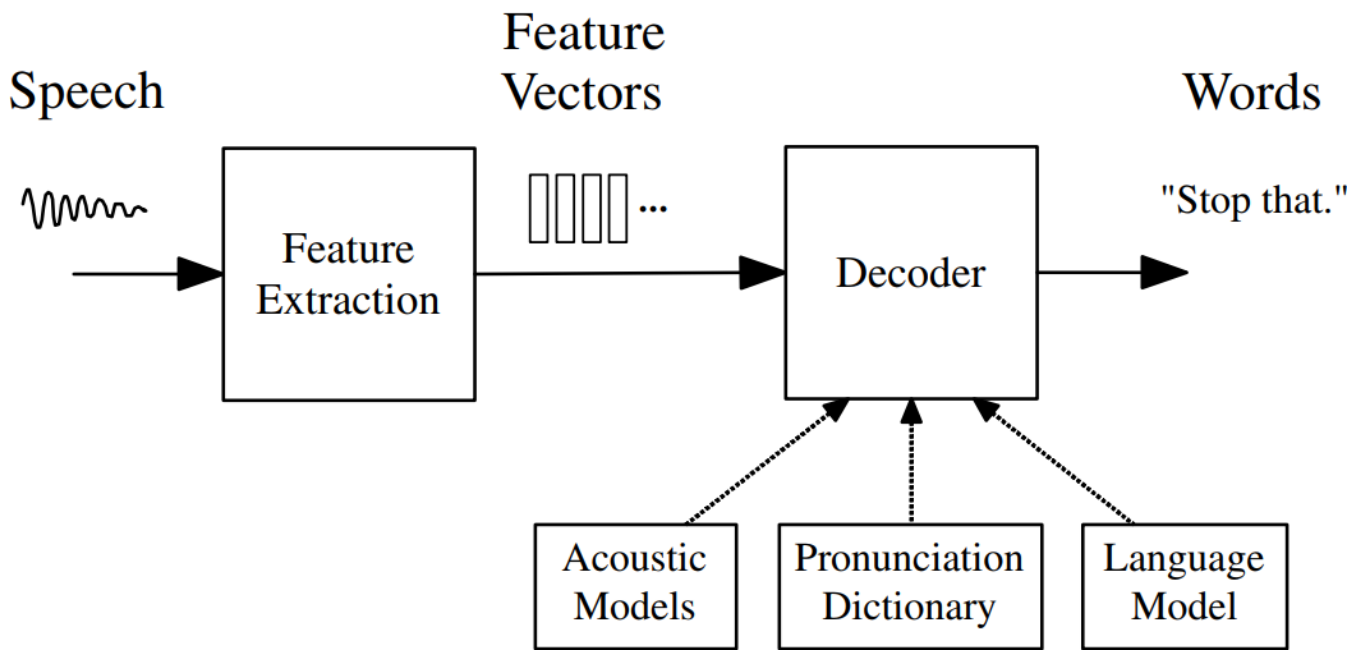
$$\hat{\mathbf{w}}_{1:L} = \arg\max_{\mathbf{w}_{1:L}} P(\mathbf{w}_{1:L}|\mathbf{y}_{1:T})$$

Since $P(\mathbf{w}_{1:L}|\mathbf{y}_{1:T})$ is difficult to model directly, Bayes' rule is the used to solve the equivalent problem

$$\hat{\mathbf{w}}_{1:L} = \arg\max_{\mathbf{w}_{1:L}} p(\mathbf{y}_{1:T}|\mathbf{w}_{1:L})P(\mathbf{w}_{1:L}),$$

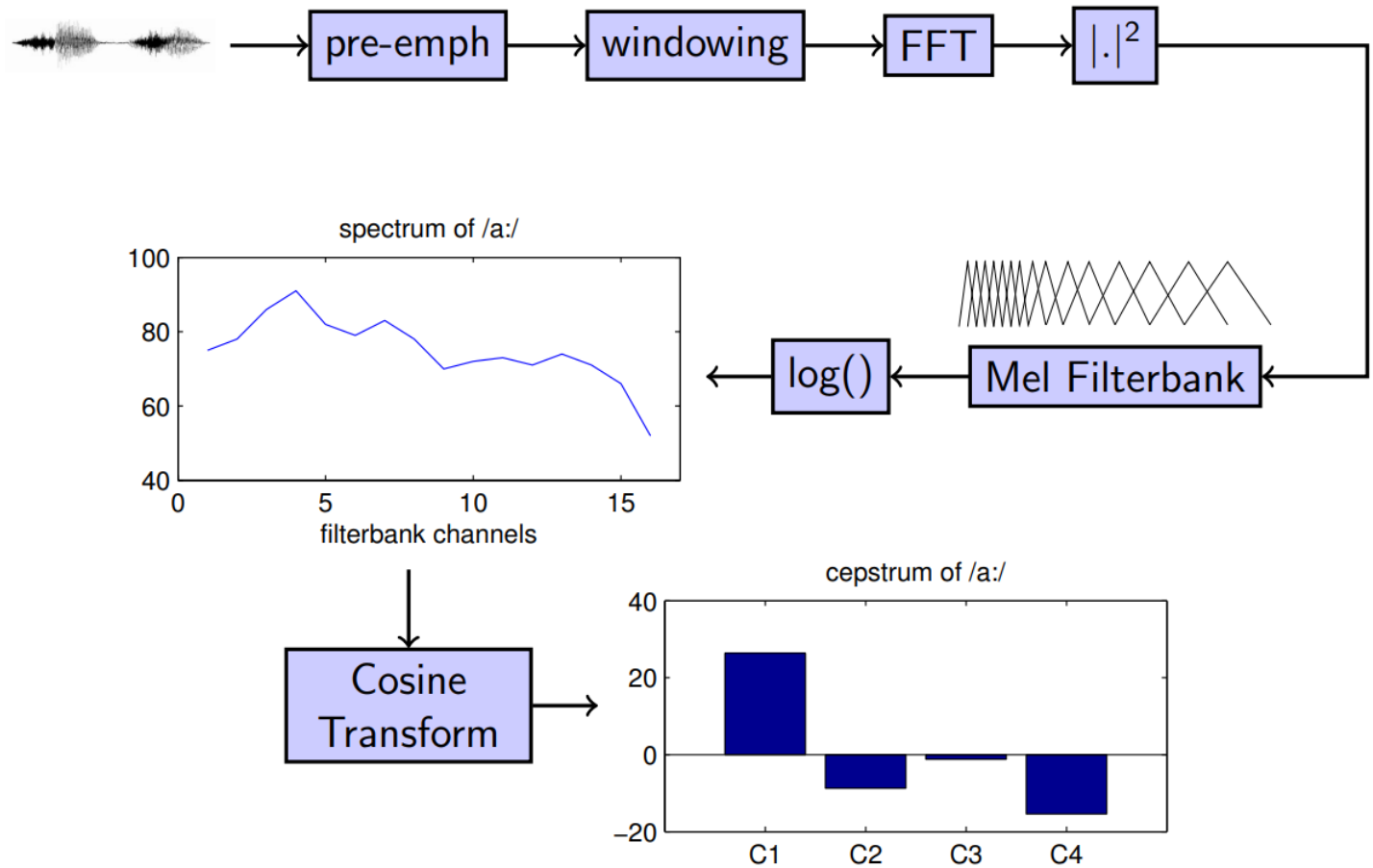
where

- the likelihood $p(\mathbf{y}_{1:T}|\mathbf{w}_{1:L})$ is the acoustic model;
- the prior $P(\mathbf{w}_{1:L})$ is the language model.


Speech
Feature Vectors
Words
Feature Extraction
...
Decoder
"Stop that."
Acoustic Models
Pronunciation Dictionary
Language Model

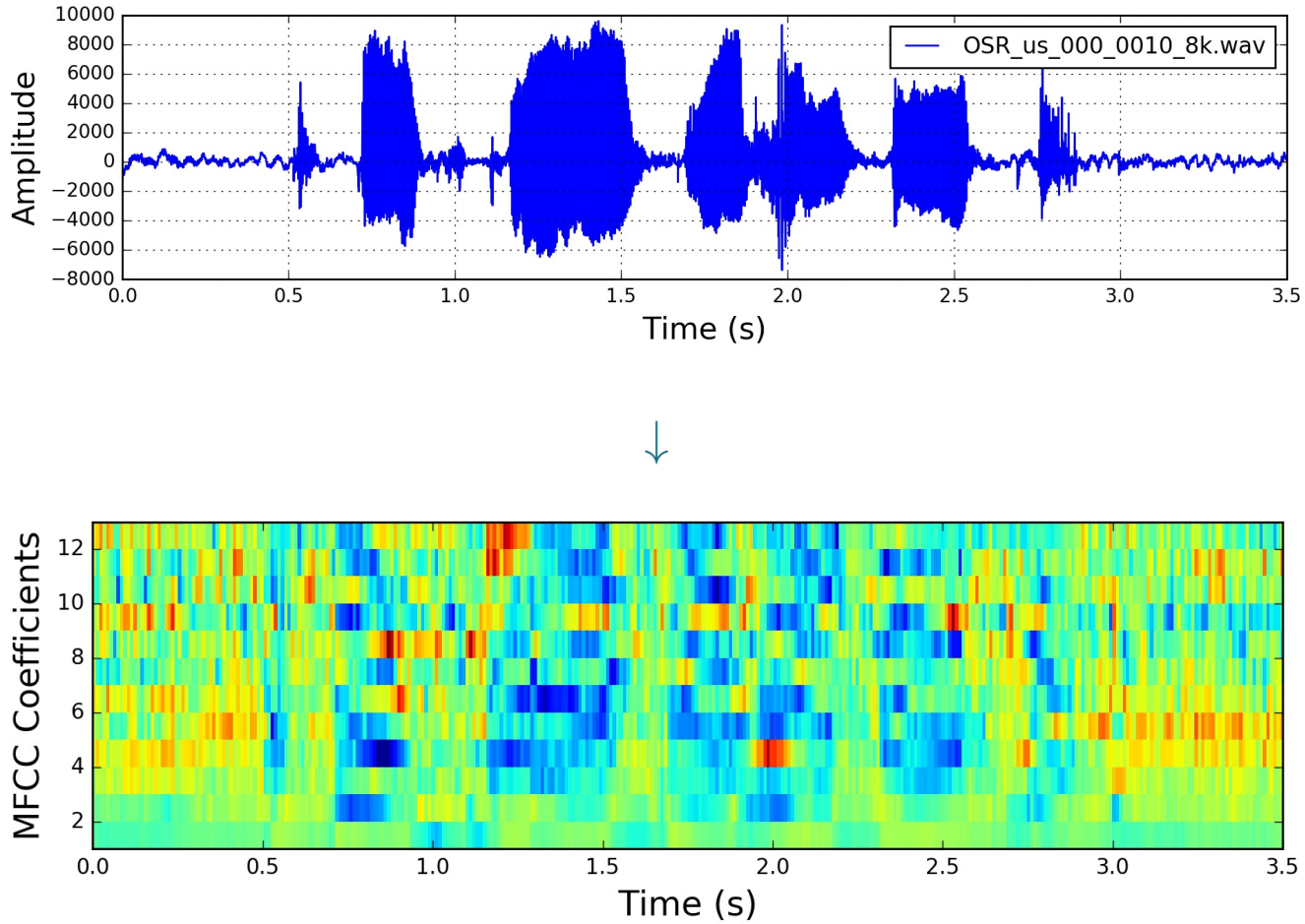
Feature extraction

- The feature extraction seeks to provide a compact representation $\mathbf{y}_{1:T}$ of the speech waveform.
- This form should minimize the loss of information that discriminates between words.
- One of the most widely used encoding schemes is based on mel-frequency cepstral coefficients (MFCCs).



MFCCs calculation.

Image credits: Giampiero Salvi, 2016. DT2118.



Feature extraction from the signal in the time domain to MFCCs.

Image credits: Haytham Fayek, 2016.

Acoustic model

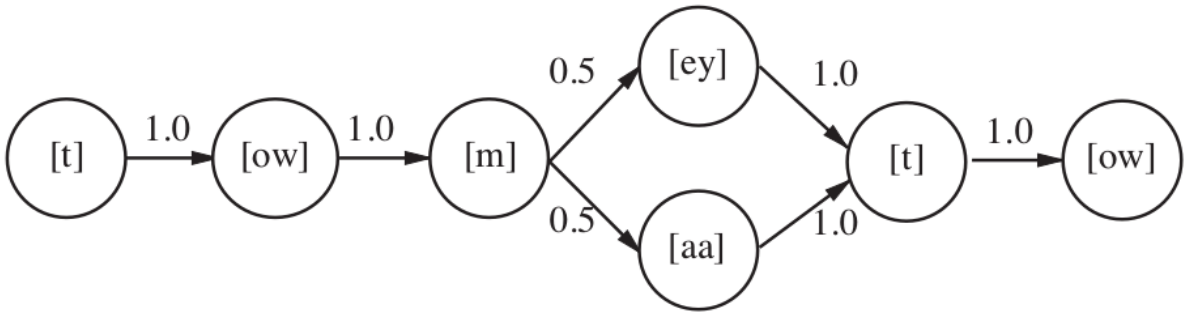
A spoken word w is decomposed into a sequence of K_w basic sounds called base phones (such as vowels or consonants).

- This sequence is called its pronunciation $\mathbf{q}_{1:K_w}^w = q_1, \ldots, q_{K_w}$.
- Pronunciations are related to words through pronunciations models defined for each word.

Vowels		Consonants B–N		Consonants P–Z	
Phone	Example	Phone	Example	Phone	Example
[iy]	b**ea**t	[b]	**b**et	[p]	**p**et
[ih]	b**i**t	[ch]	**Ch**et	[r]	**r**at
[eh]	b**e**t	[d]	**d**ebt	[s]	**s**et
[æ]	b**a**t	[f]	**f**at	[sh]	**sh**oe
[ah]	b**u**t	[g]	**g**et	[t]	**t**en
[ao]	b**ough**t	[hh]	**h**at	[th]	**th**ick
[ow]	b**oa**t	[hv]	**h**igh	[dh]	**th**at
[uh]	b**oo**k	[jh]	**j**et	[dx]	bu**tt**er
[ey]	b**ai**t	[k]	**k**ick	[v]	**v**et
[er]	B**er**t	[l]	**l**et	[w]	**w**et
[ay]	b**uy**	[el]	bott**le**	[wh]	**wh**ich
[oy]	b**oy**	[m]	**m**et	[y]	**y**et
[axr]	din**er**	[em]	bott**om**	[z]	**z**oo
[aw]	d**ow**n	[n]	**n**et	[zh]	mea**s**ure
[ax]	**a**bout	[en]	butt**on**		
[ix]	ros**e**s	[ng]	si**ng**		
[aa]	c**o**t	[eng]	wash**ing**	[-]	*silence*

Figure 23.14 The ARPA phonetic alphabet, or **ARPAbet**, listing all the phones used in American English. There are several alternative notations, including an International Phonetic Alphabet (IPA), which contains the phones in all known languages.

(a) Word model with dialect variation:



(b) Word model with coarticulation and dialect variations

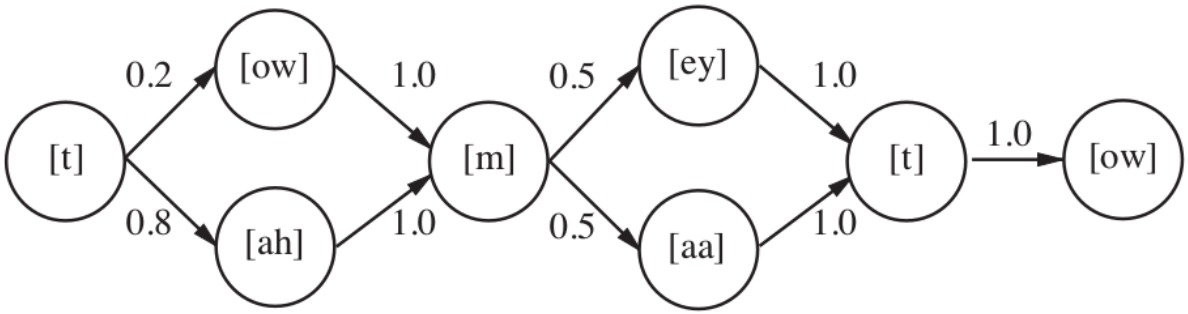
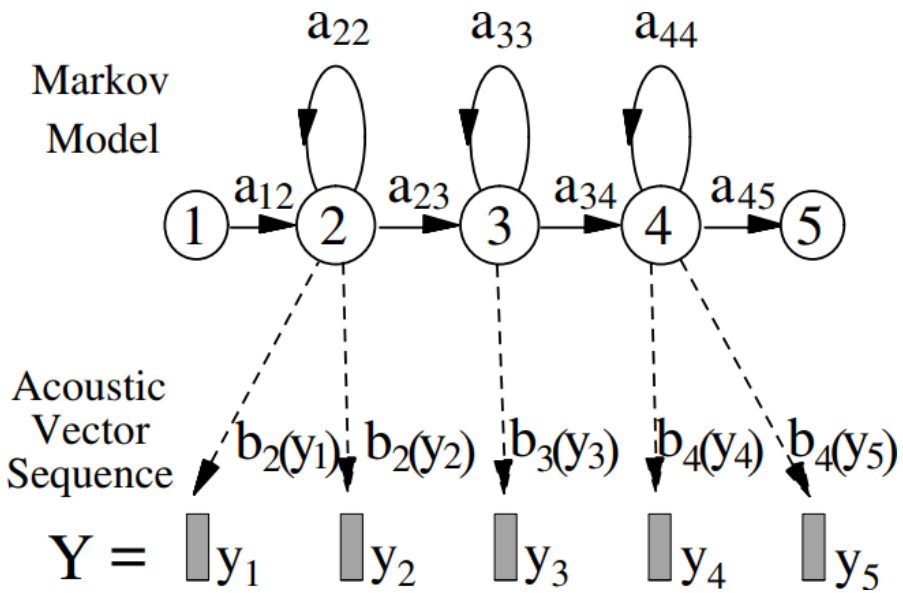


Figure 23.17 Two pronunciation models of the word "tomato." Each model is shown as a transition diagram with states as circles and arrows showing allowed transitions with their associated probabilities. (a) A model allowing for dialect differences. The 0.5 numbers are estimates based on the two authors' preferred pronunciations. (b) A model with a coarticulation effect on the first vowel, allowing either the [ow] or the [ah] phone.



Each base phone q is represented by phone model defined as a three-state continuous density HMM, where

- the transition probability parameter a_{ij} corresponds to the probability of making the particular transition from state s_i to s_j;
- the output sensor models are Gaussians $b_j(\mathbf{y}) = \mathcal{N}(\mathbf{y}; \mu^{(j)}, \Sigma^{(j)})$ and relate state variables s_j to MFCCs $\mathbf{y}$.

The full acoustic model can now be defined as a composition of pronunciation models with individual phone models:

$$p(\mathbf{y}_{1:T}|\mathbf{w}_{1:L}) = \sum_{\mathbf{Q}} P(\mathbf{y}_{1:T}|\mathbf{Q})P(\mathbf{Q}|\mathbf{w}_{1:L})$$

where the summation is over all valid pronunciation sequences for $\mathbf{w}_{1:L}$, $\mathbf{Q}$ is a particular sequence $\mathbf{q}^{w_1}, \ldots, \mathbf{q}^{w_L}$ of pronunciations,

$$P(\mathbf{Q}|\mathbf{w}_{1:L}) = \prod_{l=1}^{L} P(\mathbf{q}^{w_l}|w_l)$$

as given by the pronunciation model, and where $\mathbf{q}^{w_l}$ is a valid pronunciation for word w_l.

Given the composite HMM formed by concatenating all the constituent pronunciations $\mathbf{q}^{w_1}, \ldots, \mathbf{q}^{w_L}$ and their corresponding base phones, the acoustic likelihood is given by

$$p(\mathbf{y}_{1:T}|\mathbf{Q}) = \sum_{\mathbf{s}} p(\mathbf{s}, \mathbf{y}_{1:T}|\mathbf{Q})$$

where $\mathbf{s} = s_0, \ldots, s_{T+1}$ is a state sequence through the composite model and

$$p(\mathbf{s}, \mathbf{y}_{1:T}|\mathbf{Q}) = a_{s_0,s_1} \prod_{t=1}^{T} b_{s_t}(\mathbf{y}_t) a_{s_t s_{t+1}}.$$

From this formulation, all model parameters can be efficiently estimated from a corpus of training utterances with expectation-maximization.

N-gram language model

The prior probability of a word sequence $\mathbf{w} = w_1, ..., w_L$ is given by

$$P(\mathbf{w}) = \prod_{l=1}^{L} P(w_l | w_{l-1}, ..., w_{l-N+1}).$$

The N-gram probabilities are estimated from training texts by counting N-gram occurrences to form maximum likelihood estimates.

Decoding

The composite model corresponds to a HMM, from which the most-likely state sequence $\mathbf{s}$ can be inferred using (a variant of) Viterbi.

By construction, states $\mathbf{s}$ relate to phones, phones to pronunciations, and pronunciations to words.

Neural speech recognition

Modern speech recognition systems are now based on end-to-end deep neural network architectures trained on large corpus of data.

Deep Speech 2

- Recurrent neural network with
  - one or more convolutional input layers,
  - followed by multiple recurrent layers,
  - and one fully connected layer before a softmax layer.
- Total of 35M parameters.
- Same architecture for both English and Mandarin.

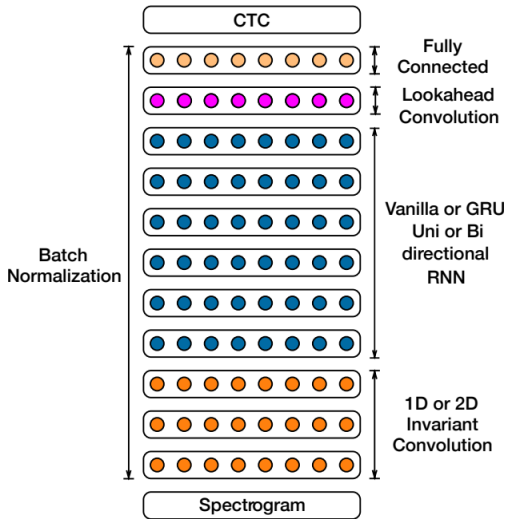


Image credits: Amodei et al, 2015. arXiv:1512.02595.

Deep Speech 2

Text-to-speech synthesis

$\mathbf{w}_{1:L}$

My name is HAL.

$\rightarrow$

$\mathbf{y}_{1:T}$

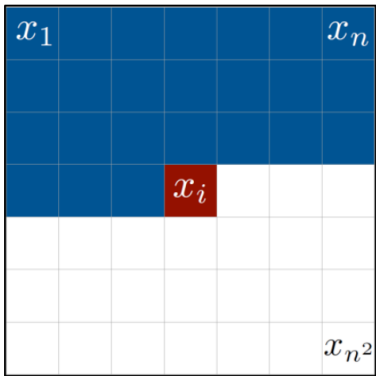
Autoregressive models

By the chain rule, any joint distribution can always be written as an incremental product of conditional distributions:

$$p(x_1, ..., x_n) = \prod_{k=1}^{n} p(x_k | x_1, ..., x_{k-1}).$$

If h_{k-1} is a lossless statistic of the previous observations in the sequence, then we can express

$$p(x_k | x_1, ..., x_{k-1}) = g_\theta(x_k | h_{k-1}).$$



Autoregressive model for images.

An autoregressive model can be formulated as a recurrent neural network such that

$$p(x_k|x_1, ..., x_{k-1}) = g_\theta(x_k|h_{k-1})$$
$$h_k = f_\theta(x_k, h_{k-1})$$

where

- g specifies a probability density function for the next x given h.
- f specifies (in a deterministic way) the next state h given x.

If g and f have enough capacity (e.g., f and g are large neural networks), then the resulting autoregressive model is often capable of modeling complex distributions, such as text, images or speech.

The same architecture can be used for modeling conditional distributions over inputs.

Tacotron 2

The Tacotron 2 system is a sequence-to-sequence neural network architecture for text-to-speech. It consists of two components:

- a recurrent sequence-to-sequence feature prediction network with attention which predicts a sequence of mel spectrogram frames from an input character sequence;
- a Wavenet vocoder which generates time-domain waveform samples conditioned on the predicted mel spectrogram frames.

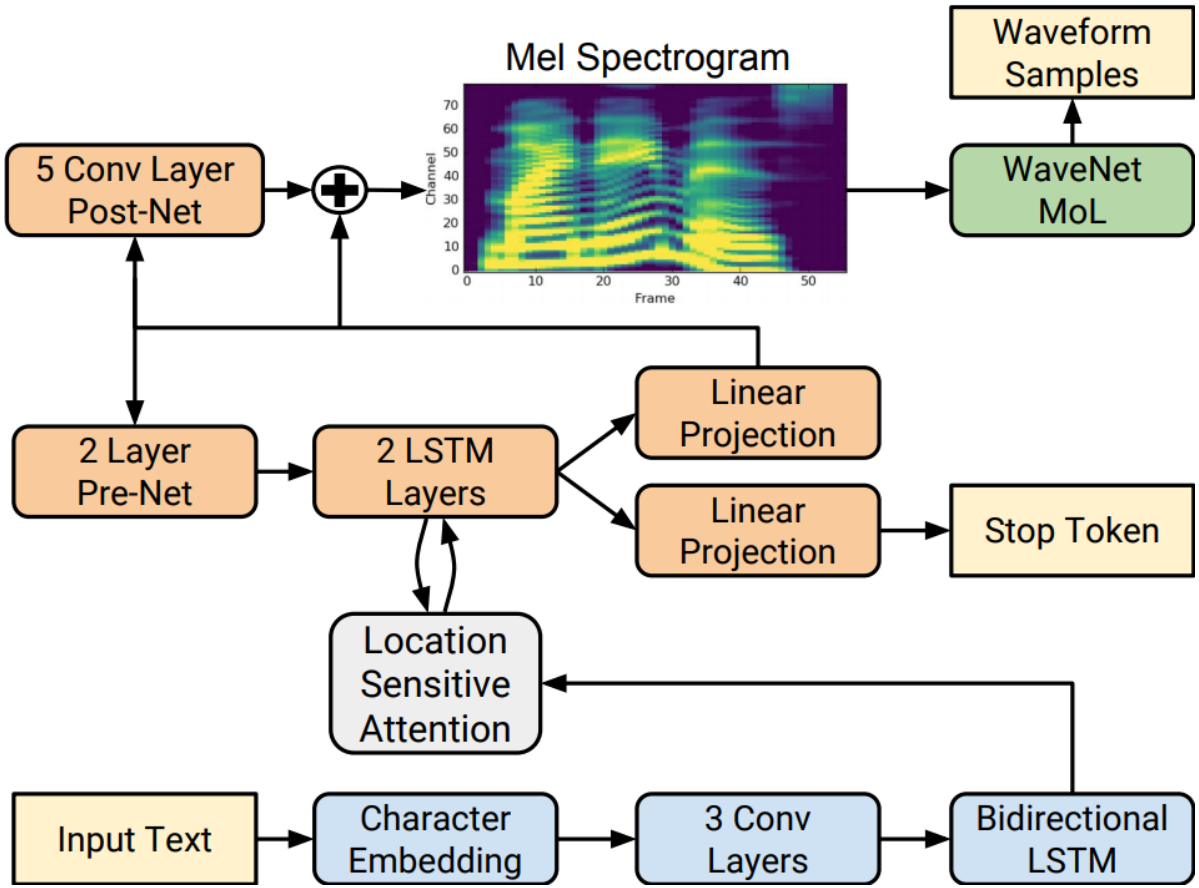
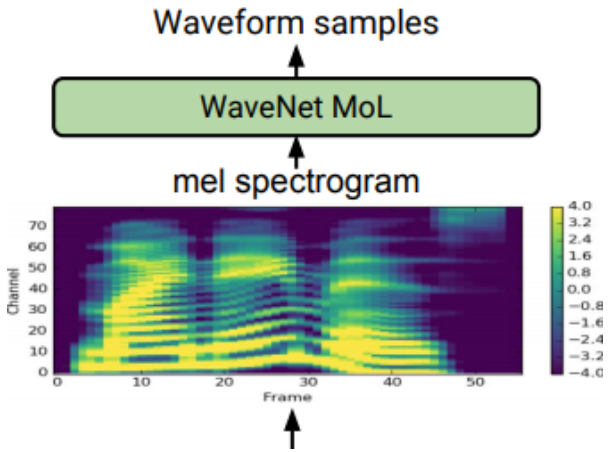
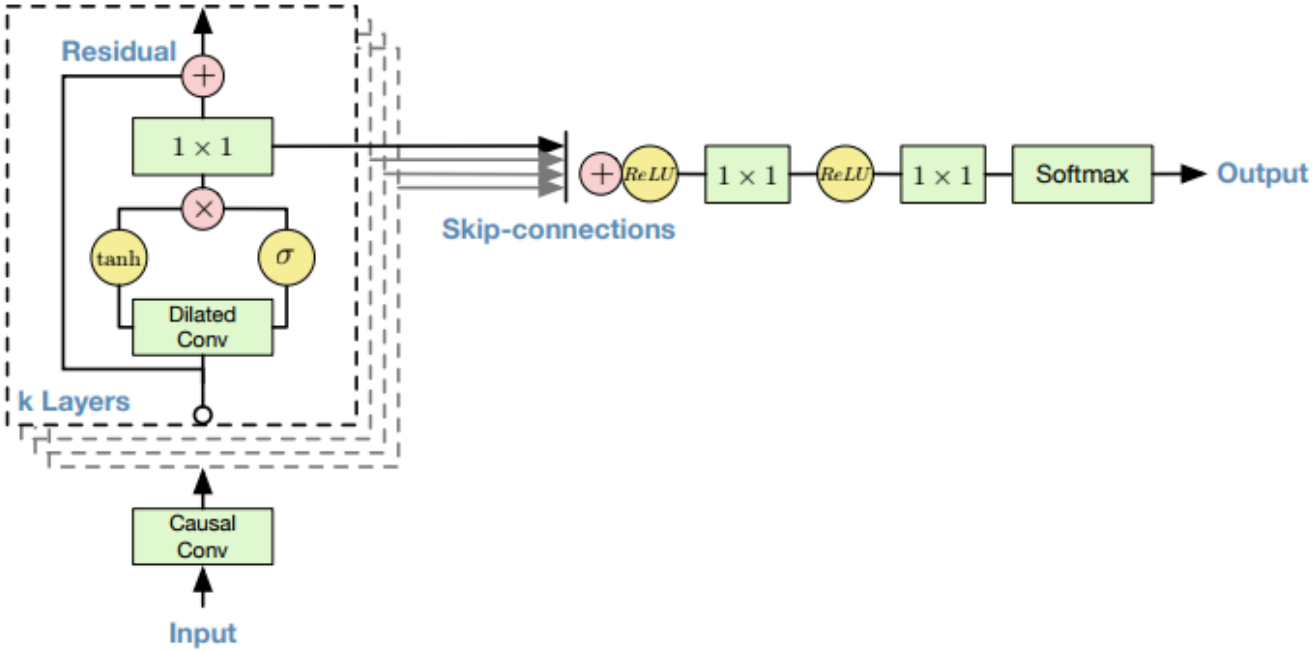


Image credits: Shen et al, 2017. arXiv:1712.05884.

Wavenet

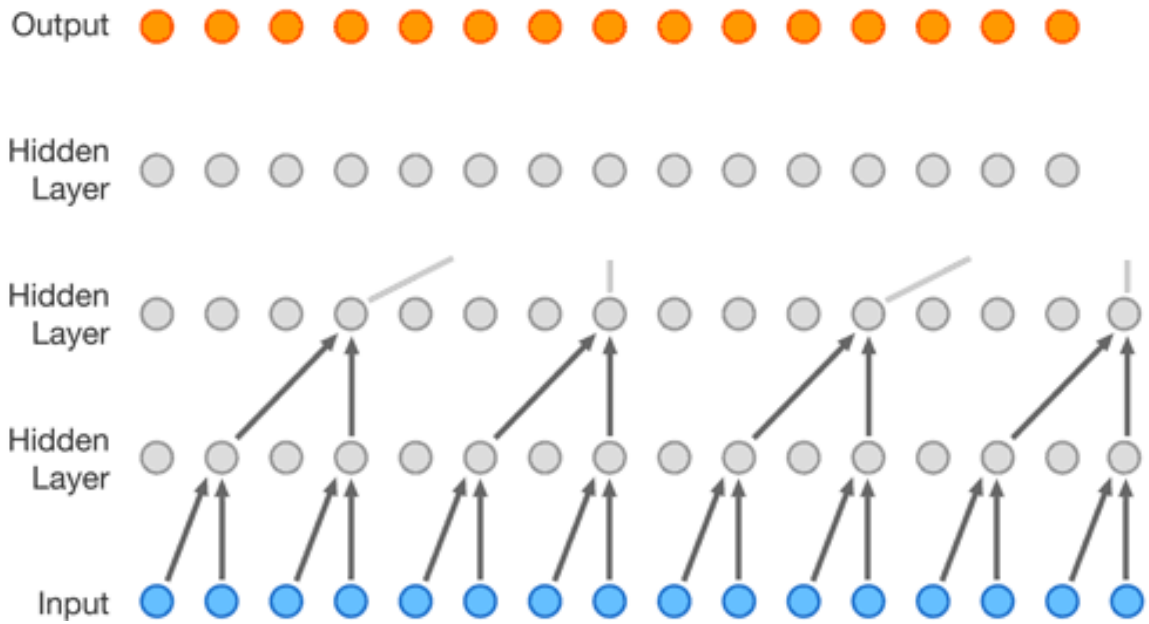
- The Tacotron 2 architecture produces mel spectrograms as outputs, which remain to be synthesized as waveforms.
- This last step can be performed through another autoregressive neural model, such as Wavenet, to transform mel-scale spectrograms into high-fidelity waveforms.





The Wavenet architecture.

Image credits: Van den Oord et al, 2016..



Dilated convolutions.

Image credits: Van den Oord et al, 2016..

Audio samples at

- deepmind.com/blog/wavenet-generative-model-raw-audio
- google.github.io/tacotron

Google Assistant: Soon in your smartphone.

Summary

- Natural language understanding is one of the most important subfields of AI.
- Machine translation, speech recognition and text-to-speech synthesis are instances of sequence-to-sequence problems.
- All problems can be tackled with traditional statistical inference methods but require sophisticated engineering.
- State-of-the-art methods are now based on neural networks.

The end.

References

- Gales, M., & Young, S. (2008). The application of hidden Markov models in speech recognition. Foundations and Trends® in Signal Processing, 1(3), 195-304.