Modeling in Automatic Speech Recognition: Beyond Hidden Markov Models

Ralf Schlüter, P. Bahar, E. Beck, P. Dötsch, K. Irie, C. Lüscher, A. Merboldt, Z. Tüske, P. Golik, A. Zeyer, H. Ney.

Human Language Technology and Pattern Recognition
Lehrstuhl Informatik 6
Fakultät für Mathematik, Informatik und Naturwissenschaften
RWTH Aachen University

Outline

Introduction

Current situation in Automatic Speech Recognition (ASR):

- Decade brought **≳50% relative improvements in WER** by introducing artifical neural networks to all levels of modeling.
- Traditional state-of-the-art challenged by novel "end-to-end" ASR architectures.
- Enabling factor: generic machine learning tools, developed for diverse and complex tasks.

ASR very challenging task - advantages from a method evaluation viewpoint:

- Provides clear performance objective.
- Strong state-of-the-art performance to compete against for new approaches.
- Various and diverse well-covered benchmarks.

Topics of interest:

- performance vs. system complexity
- variable length sequence alignment: beyond HMM
- primary training data and secondary knowledge sources
- reusability of inferred knowledge

Outline

Sequence Classification

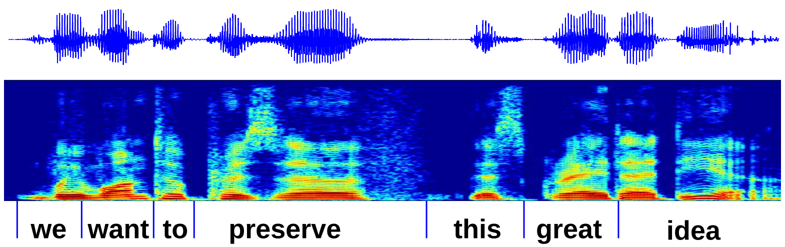
Tasks:

- automatic speech recognition
- text image recognition
- machine translation

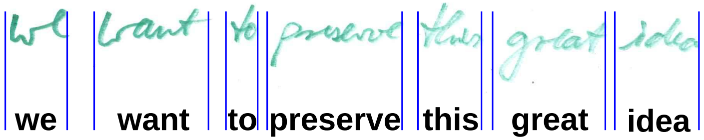
Most **general case**:

- input sequence:
  $X := x_1 \ldots x_t \ldots x_T$
- output sequence (of unknown length N):
  $W := w_1 \ldots w_n \ldots w_N$
- true distribution $pr(W|X)$
  (can be extremely complex!)

Speech Recognition



Text Image Recognition



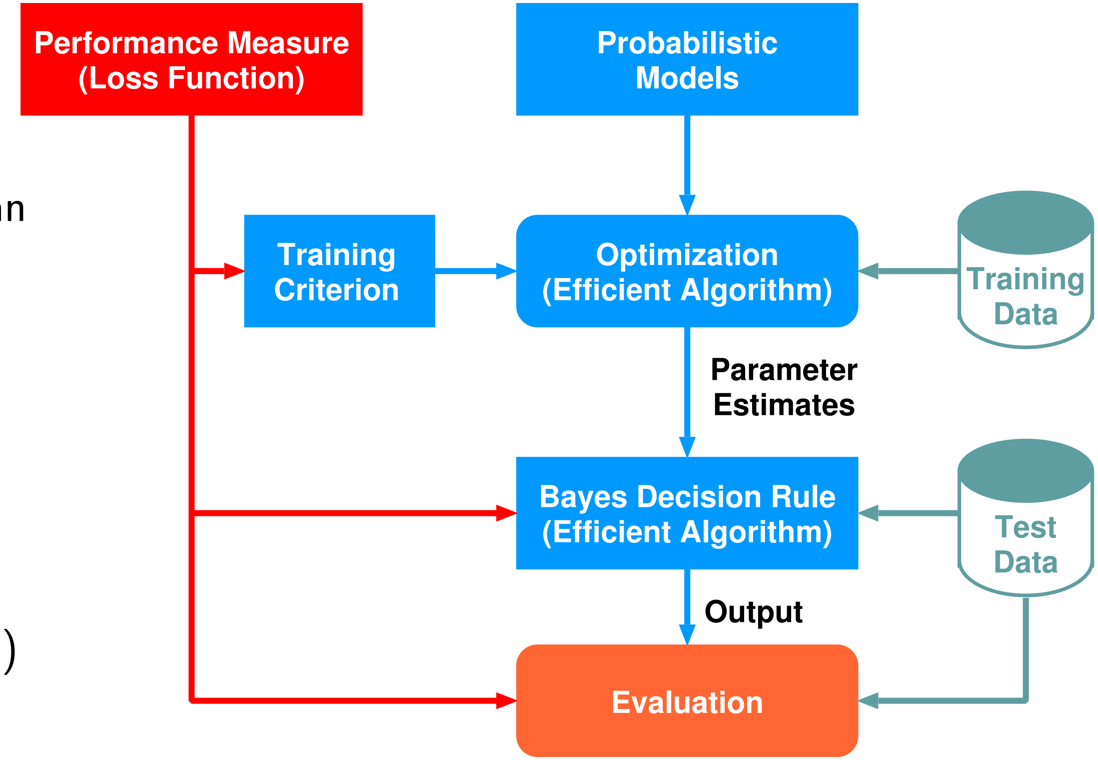
Machine Translation

wir wollen diese große Idee erhalten

we want to preserve this great idea

Statistical Approach Revisited

- **Performance measure**:
  judges quality of system output
- **Probabilistic models**:
  capture dependencies
  - elementary observations: Gaussian mixture, log-linear, SVM, NN, ...
  - strings: *n*-gram Markov chain, HMM, CRF, RNN, LSTM, attention/transformer, ...
- **Training criterion**:
  learns free parameters of models
  - linked to performance criterion?
  - complex optimization (efficiency!)
- **Bayes decision rule**:
  generates output word sequence
  - combinatorial problem (efficient algorithms: dynamic programming, beam search, A*, ...



Speech Recognition = Modeling + Statistics + Efficient Algorithms

Sequence Decision Rule

- performance measure or **loss function** $L[\widetilde{w}_1^{\widetilde{N}}, w_1^N]$ (e.g. edit distance for word/phoneme/character error computation) between true output sequence $\widetilde{w}_1^{\widetilde{N}}$ and hypothesized output sequence w_1^N.
- **Bayes decision rule** minimizes expected loss:

$$x_1^T \to r_L(x_1^T) := \arg\min_{w_1^N} \left\{ \sum_{\widetilde{w}_1^{\widetilde{N}}} pr(\widetilde{w}_1^{\widetilde{N}}|x_1^T) \cdot L[\widetilde{w}_1^{\widetilde{N}}, w_1^N] \right\}$$

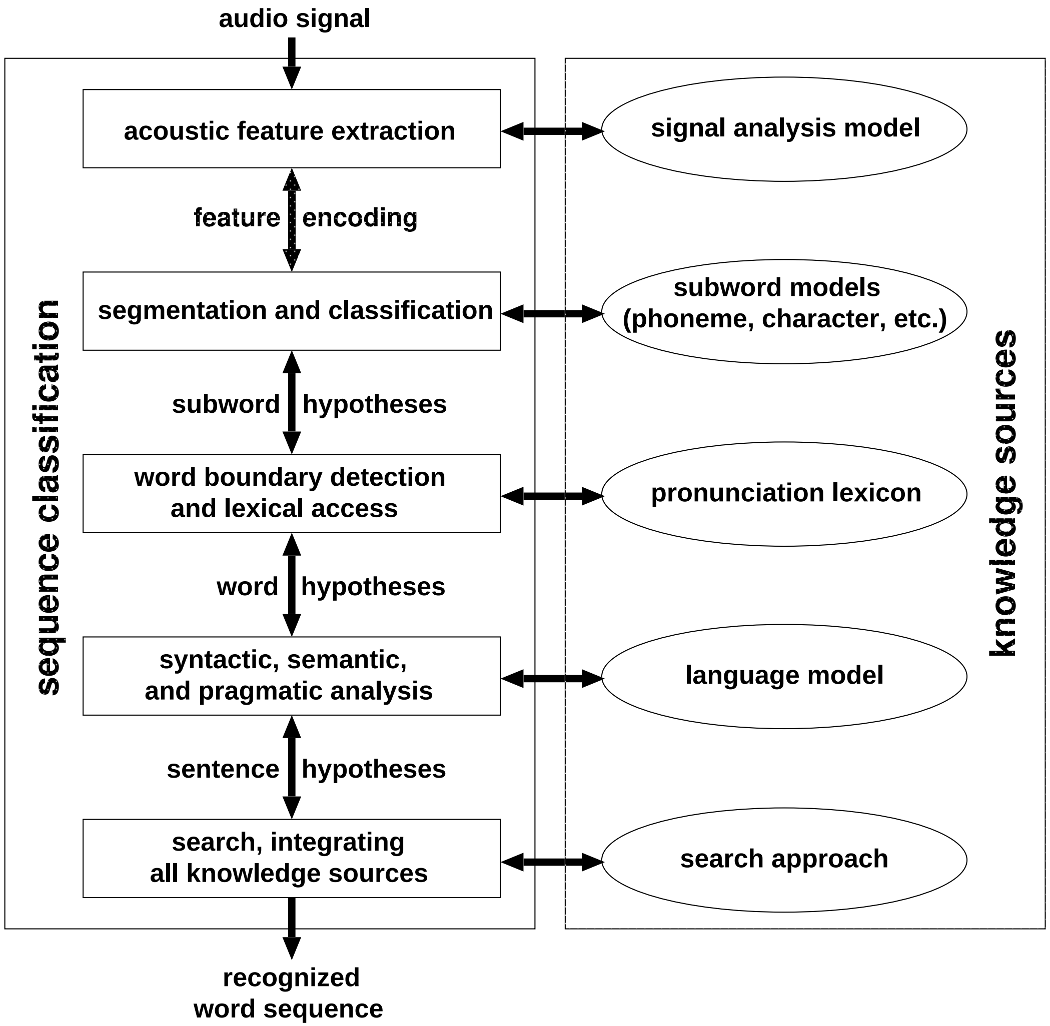
- Standard decision rule uses sequence-level **zero-one loss**: minimizes **sentence error**

$$x_1^T \to r_{0\text{-}1}(x_1^T) := \arg\max_{w_1^N} \left\{ pr(w_1^N|x_1^T) \right\}$$

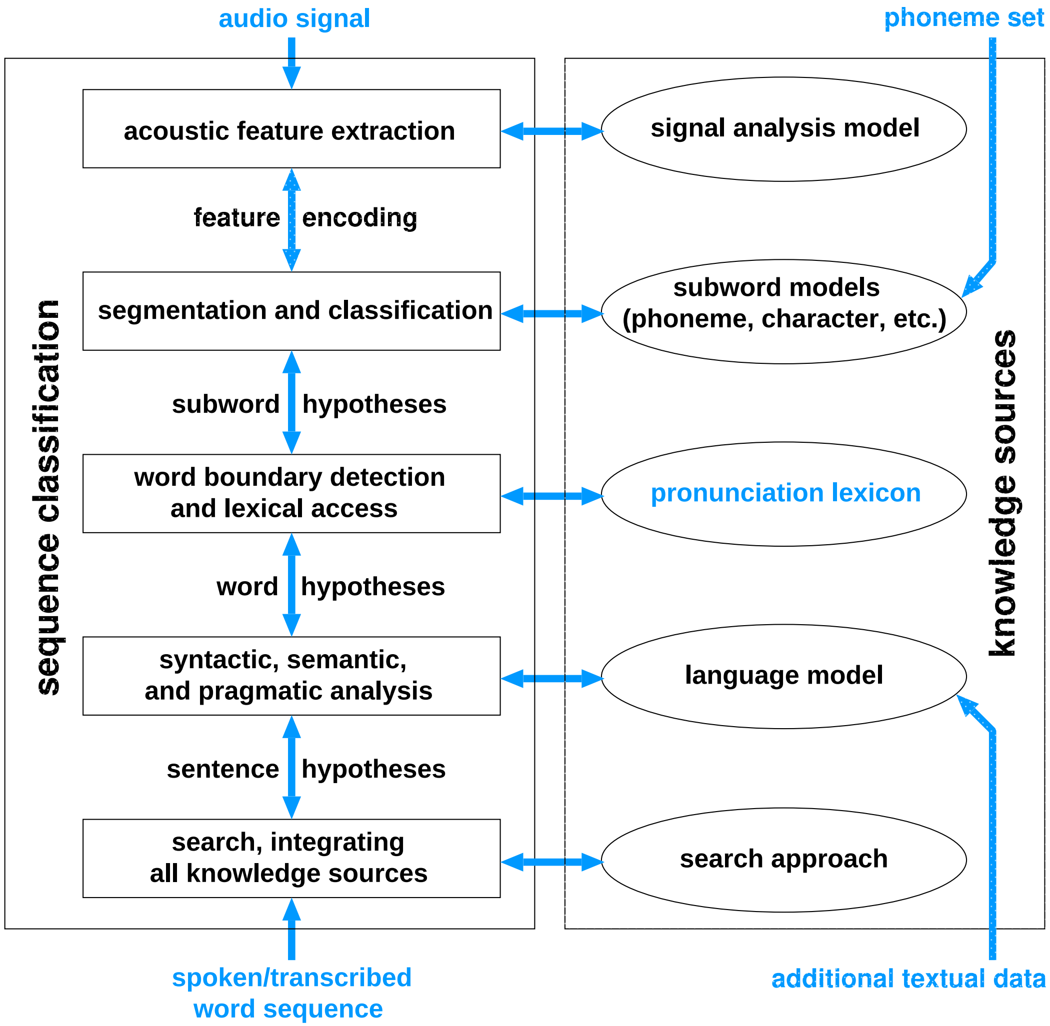
  Since [Bahl & Jelinek$^+$ 1983], this simpified Bayes decision rule is widely used for speech recognition, handwriting recognition, machine translation, ...
- Works well, as often both decision rules **coincide**.
  This can be proven under certain conditions [Schlüter & Nussbaum$^+$ 2012], e.g.:

$$L[w_1^N, \widetilde{w}_1^{\widetilde{N}}] \text{ is a metric, and} \quad \max_{w_1^N} pr(w_1^N|x_1^T) \geq 0.5 \quad \Rightarrow \quad r_L(x_1^T) = r_{0\text{-}1}(x_1^T)$$

Statistical Approach: Integrated Decisions End-to-End



Statistical Approach: Training End-To-End



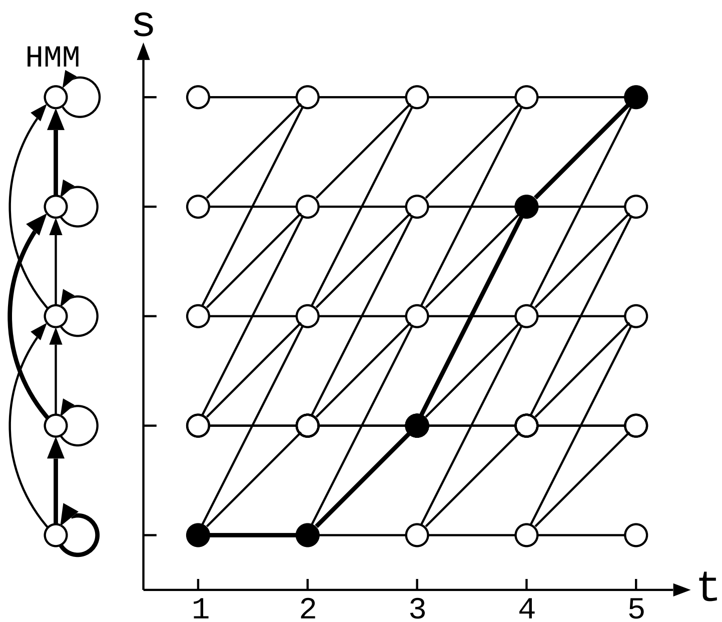
Outline

Sequence Modeling

- Problem in Bayes decision rule:
  - true posterior distribution: unknown
  - separation into language model and acoustic model

$$p(w_1^N|x_1^T) = \frac{p(w_1^N) \cdot p(x_1^T|w_1^N)}{p(x_1^T)}$$

- Acoustic model $p(x_1^T|w_1^N)$: links sentence hypothesis w_1^N to observation sequence x_1^T.
- Problem in ASR: speaking rate variation
  $\rightarrow$ non-linear time alignment
- **Hidden Markov model**:
  - linear chain of states $s = 1, ..., S$
  - transitions: forward, loop and skip
  - emissions: Gaussian mixture distributions (originally)

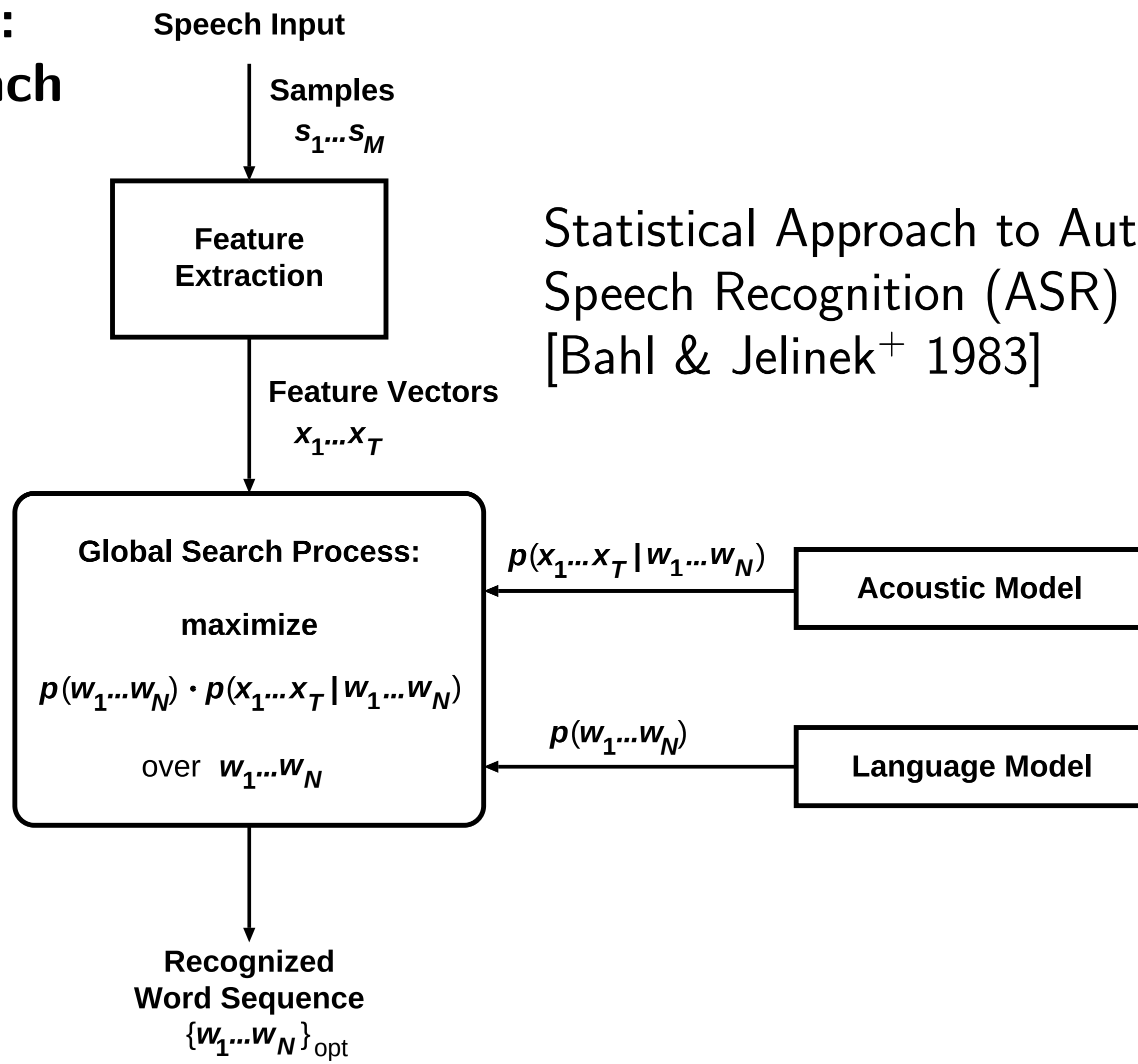


- Trellis:
  - unfold HMM over time $t = 1, ..., T$
  - path: state sequence $s_1^T = s_1 \ldots s_t \ldots s_T$
  - observations: $x_1^T = x_1 \ldots x_t \ldots x_T$

- **acoustic model** using hidden state sequences s_1^T:

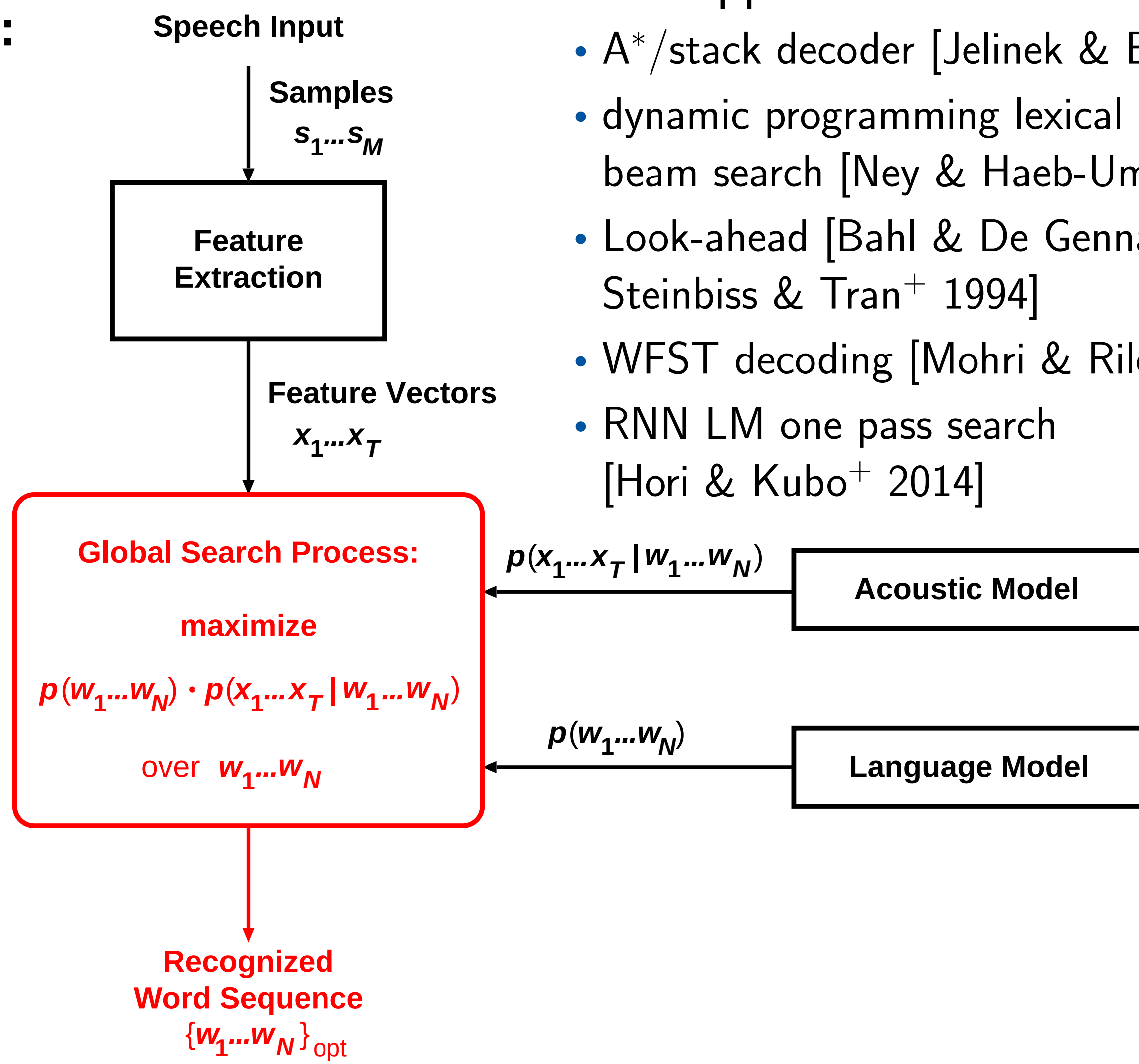
$$p(x_1^T|w_1^N) = \sum_{s_1^T} p(x_1^T, s_1^T|w_1^N) = \sum_{s_1^T} \prod_t [\underbrace{p(s_t|s_{t-1}, w_1^N)}_{\text{transition}} \cdot \underbrace{p(x_t|s_t, w_1^N)}_{\text{emission}}]$$

ASR Architecture: Statistical Approach



Statistical Approach to Automatic Speech Recognition (ASR) [Bahl & Jelinek+ 1983]

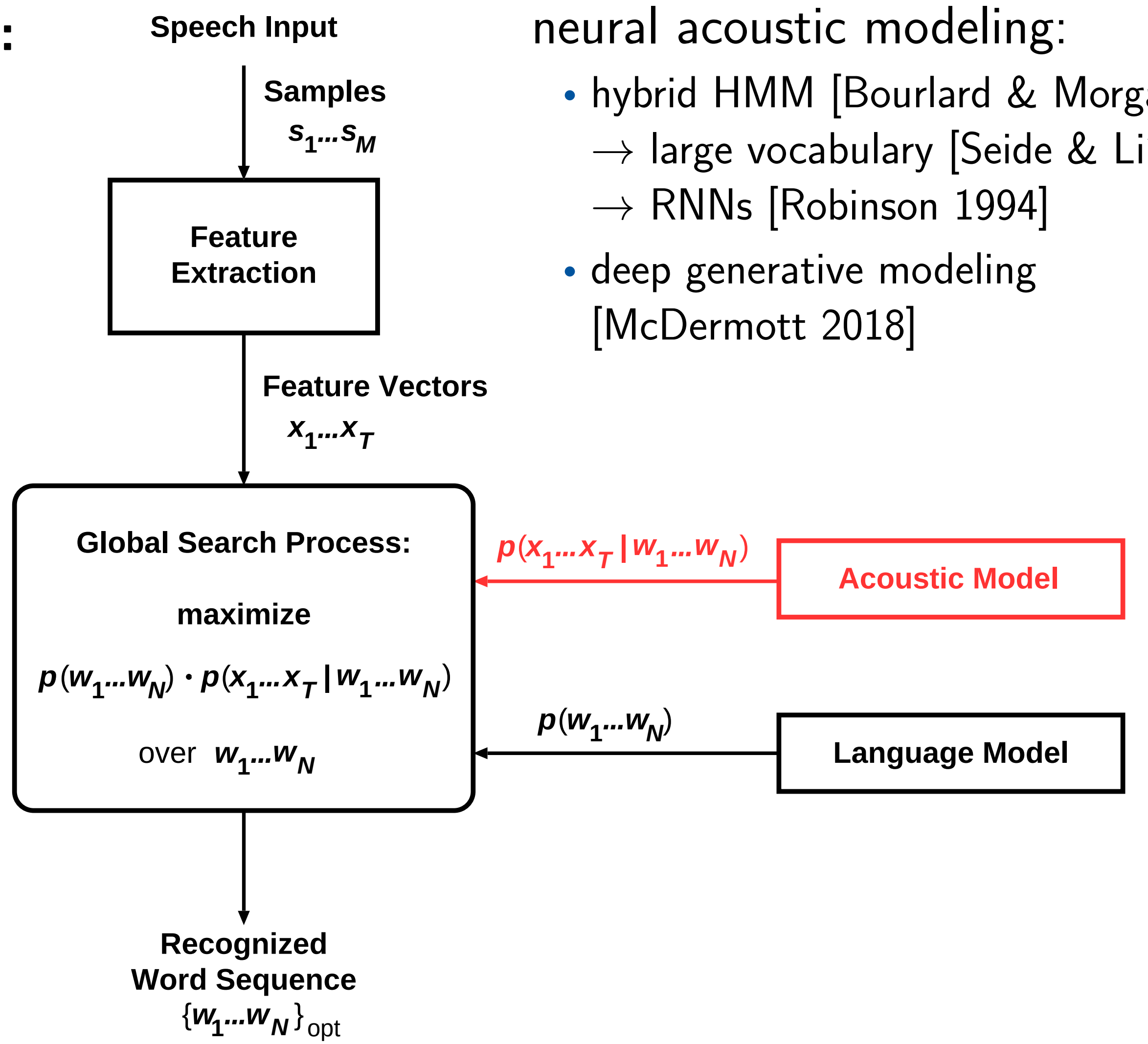
ASR Architecture: Search



search approaches:

- A*/stack decoder [Jelinek & Bahl+ 1975]
- dynamic programming lexical prefix tree beam search [Ney & Haeb-Umbach+ 1992]
- Look-ahead [Bahl & De Gennaro+ 1993, Steinbiss & Tran+ 1994]
- WFST decoding [Mohri & Riley 1999]
- RNN LM one pass search [Hori & Kubo+ 2014]

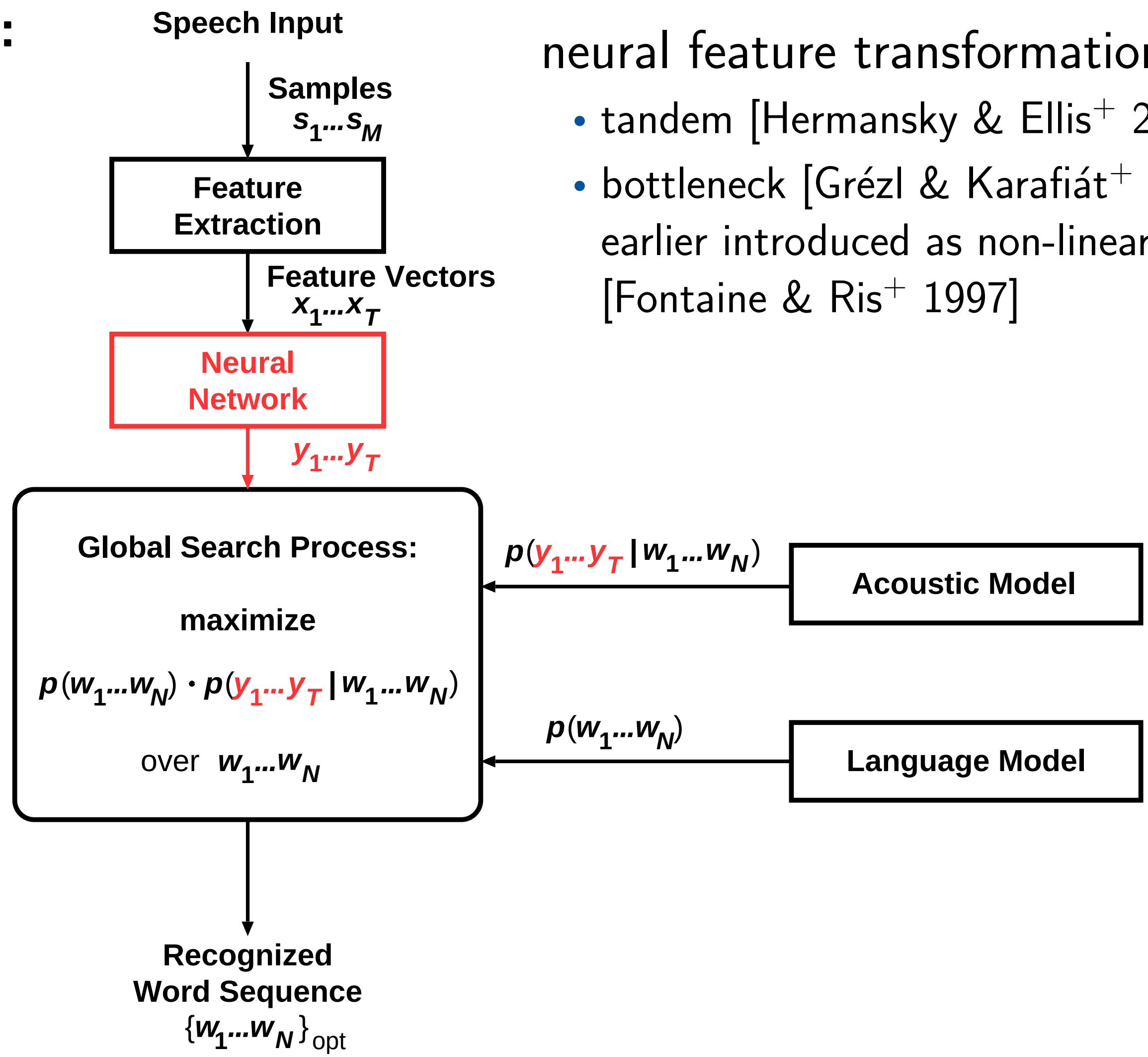
ASR Architecture: Neural Networks



neural acoustic modeling:

- hybrid HMM [Bourlard & Morgan 1993]
  $\rightarrow$ large vocabulary [Seide & Li$^+$ 2011]
  $\rightarrow$ RNNs [Robinson 1994]
- deep generative modeling [McDermott 2018]

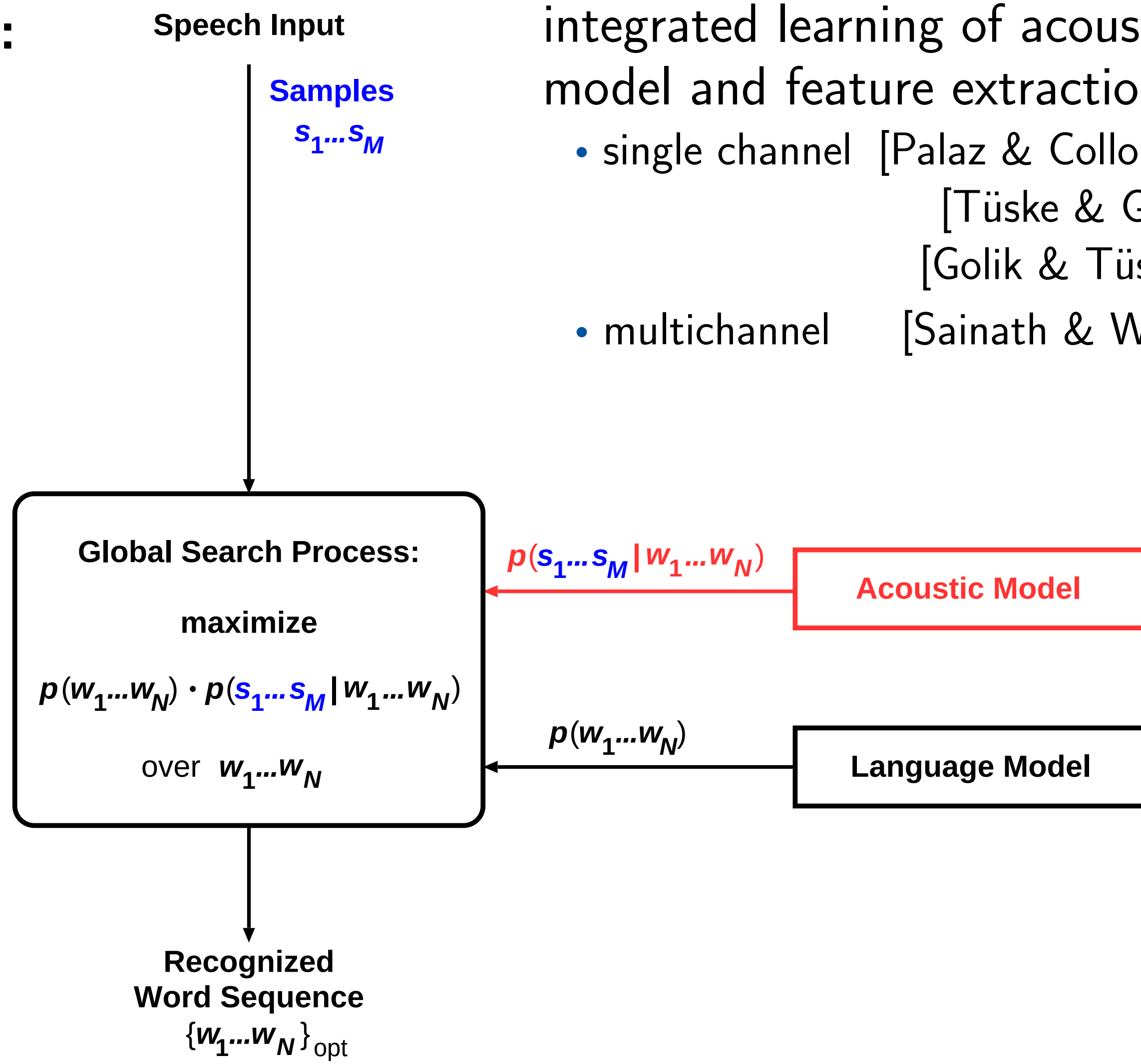
ASR Architecture: Neural Networks



neural feature transformation:

- tandem [Hermansky & Ellis+ 2000]
- bottleneck [Grézl & Karafiát+ 2007] earlier introduced as non-linear LDA [Fontaine & Ris+ 1997]

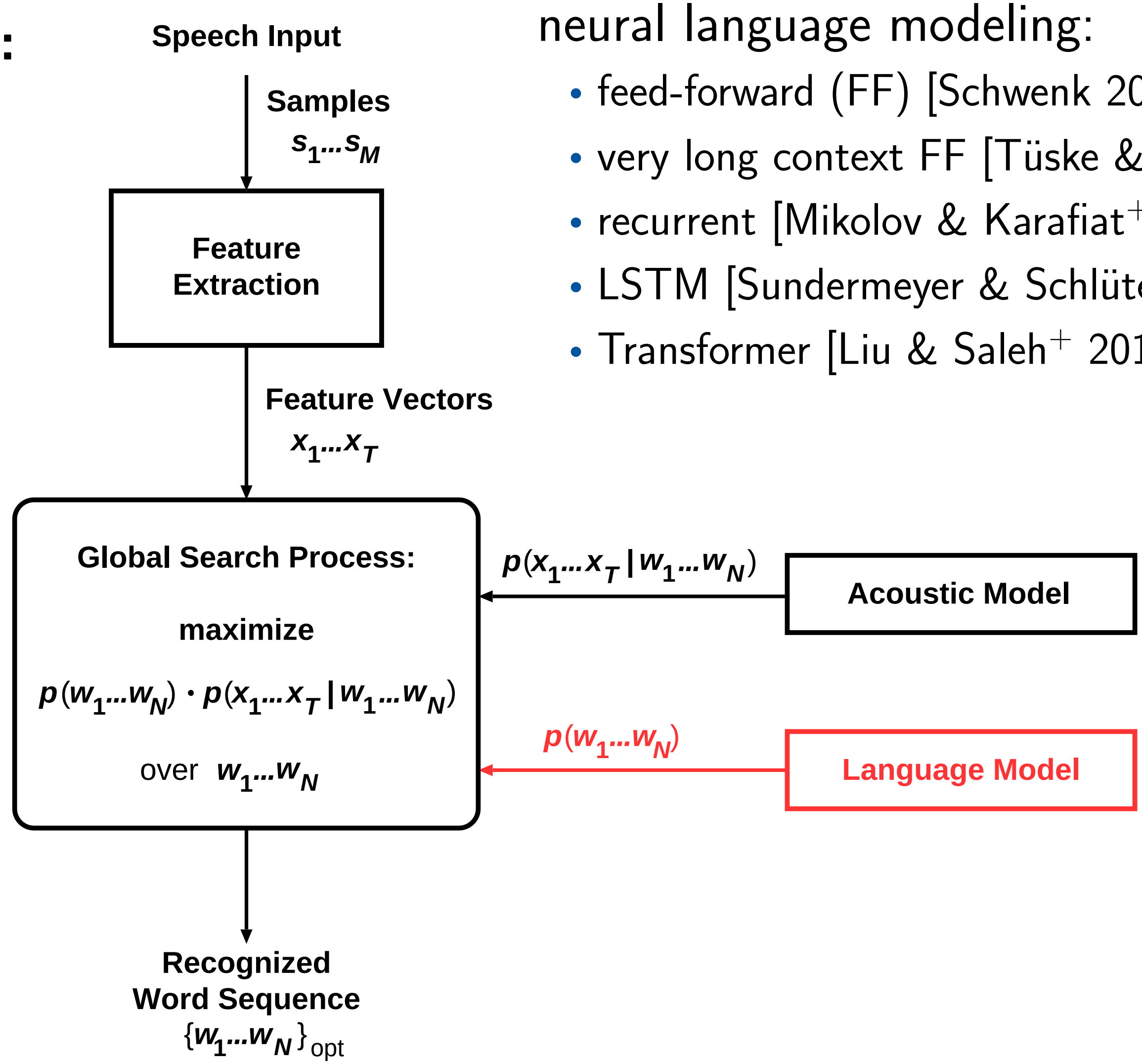
ASR Architecture: Neural Networks



integrated learning of acoustic model and feature extraction

- single channel [Palaz & Collobert+ 2013]
  [Tüske & Golik+ 2014]
  [Golik & Tüske+ 2015a]
- multichannel [Sainath & Weiss+ 2015]

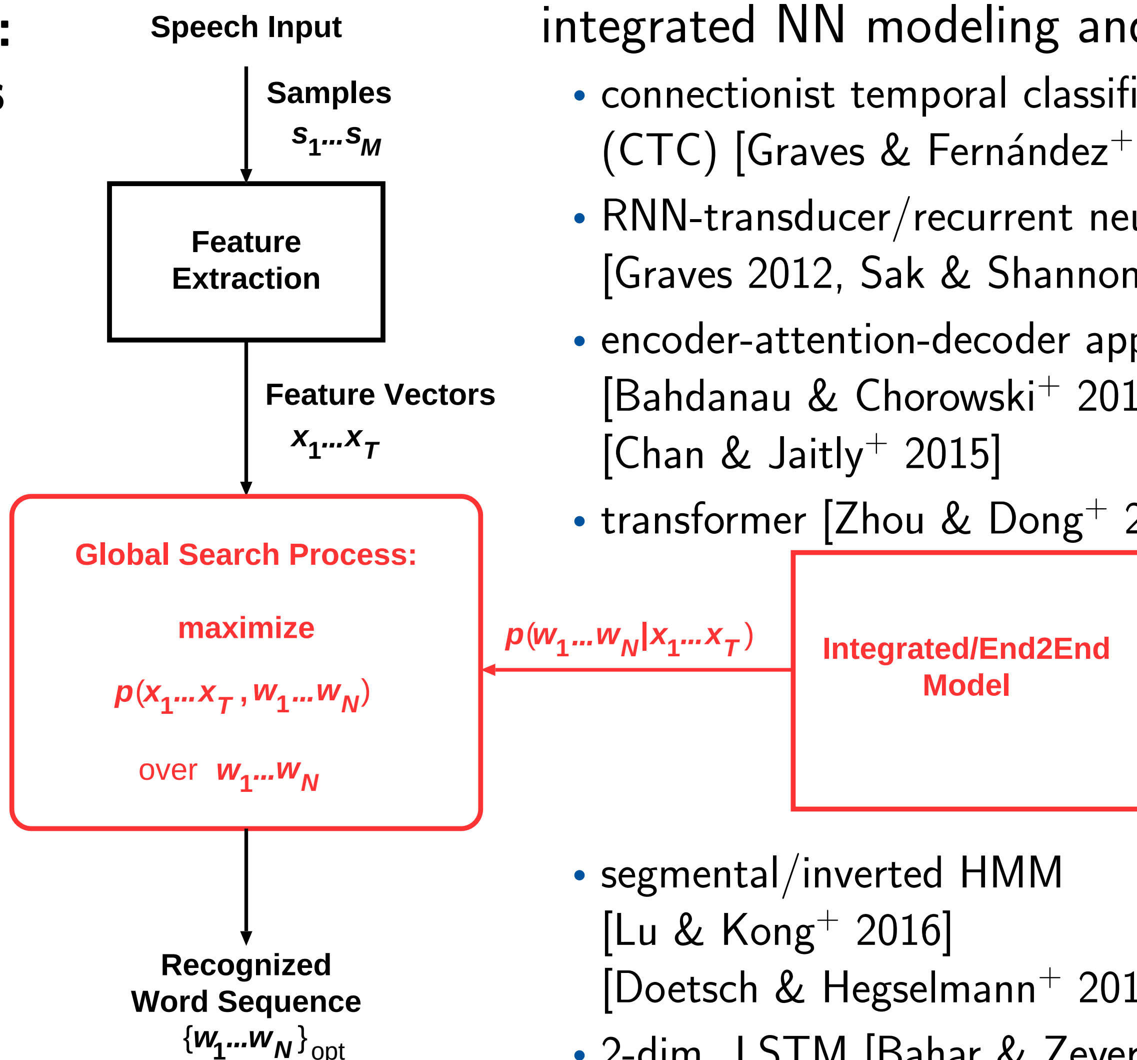
ASR Architecture: Neural Networks



neural language modeling:

- feed-forward (FF) [Schwenk 2007]
- very long context FF [Tüske & Irie+ 2016]
- recurrent [Mikolov & Karafiat+ 2010]
- LSTM [Sundermeyer & Schlüter+ 2012]
- Transformer [Liu & Saleh+ 2018]

ASR Architecture: Novel Approaches



integrated NN modeling and search:

- connectionist temporal classification (CTC) [Graves & Fernández+ 2006]
- RNN-transducer/recurrent neural aligner [Graves 2012, Sak & Shannon+ 2017]
- encoder-attention-decoder approach [Bahdanau & Chorowski+ 2015] [Chan & Jaitly+ 2015]
- transformer [Zhou & Dong+ 2018]
- segmental/inverted HMM [Lu & Kong+ 2016] [Doetsch & Hegselmann+ 2016]
- 2-dim. LSTM [Bahar & Zeyer+ 2019]

Outline

Outline

Neural Network Acoustic Modeling within Standard HMM Approach

- Decomposition within **Bayes decision rule** [Bahl & Jelinek+ 1983]:

$$\underset{w_1^N}{\operatorname{argmax}}\, p(w_1^N|x_1^T) = \underset{w_1^N}{\operatorname{argmax}}\, p(w_1^N) \cdot p(x_1^T|w_1^N)$$

- Decomposition of first order **HMM**:

$$p(x_1^T|w_1^N) = \sum_{s_1^T} \prod_{t=1}^{T} p(x_t|s_t, w_1^N) \cdot p(s_t|s_{t-1}, w_1^N)$$

- emission probability distribution using **Gaussian mixture**:
  - Gaussian mixture distribution:

$$p(x_t|s_t, w_1^N) = \sum_{l} c_{s_t l} \mathcal{N}(x_t|\mu_{s_t l}, \Sigma_{sl})$$

  - (state) posterior level:
    + Gaussian w/pooled covariance equivalent to log-linear model with linear features
    + Gaussian mixture equivalent to log-linear mixture model
- Possibilities to introduce neural network modeling while keeping the **HMM alignment** process?

Tandem [Hermansky & Ellis+ 2000]

- **Idea**: use (properly transformed) ANN outputs to augment acoustic feature set
- First ANN-approach to considerably improve LVCSR on top of Gaussian-mixture HMMs
- **Approach**:
  - train phone-classifier ANN, use its output, or the output of intermediate/hidden layers as (additional) features for Gaussian mixture HMMs,
  - variant: **bottleneck** features [Grézl & Karafiát+ 2007], earlier introduced as non-linear discriminant analysis [Fontaine & Ris+ 1997]
  - usually requires less labels for NN training, than hybrid DNN/HMM approach.
  - Typically, some post-processing is applied to the neural network output: log, decorrelation and dimension reduction with PCA, concatenation with basic acoustic feature set (e.g. MFCC).
- **Advantages**:
  - all techniques from Gaussian mixture HMM modeling can be used, in particular speaker adaptation and discriminative (sequence) training
  - cross-/multi-lingual training data exploitable [Stolcke & Grézl+ 2006, Tüske & Pinto+ 2013]
  - bootstrapping on minimal amounts of target task training data [Golik & Tüske+ 2015b]
- **Disadvantage**: training usually twofold and thus inconsistent - two models required: tandem DNN and Gaussian mixture or hybrid HMM (yet: fine-tuning end-to-end [Tüske & Golik+ 2015])

Hybrid HMM: modeling the acoustic vector x_t [Bourlard & Morgan 1993]

- Phonetic labels (allophones, sub-phones): $(s, w_1^N) \rightarrow \phi = \phi_{s,w_1^N}$
- Typical approach: decision trees, e.g. classification and regression trees (CART):
- **Hidden Markov model (HMM)** emission probability density:

$$p(x_t|s, w_1^N) = p(x_t|\phi_{s,w_1^N})$$

- **Idea**: rewrite the emission probability for label ϕ and acoustic vector x_t:

$$p(x_t|\phi) = \frac{p(x_t) \cdot p(\phi|x_t)}{p(\phi)}$$

  – prior probability $p(\phi)$: estimated as relative frequencies (alternatively averaged NN posteriors)
  – for recognition purposes: term $p(x_t)$ can be dropped
- **Result**: rather than the phone label emission distribution $p(x_t|\phi)$, model the phone label posterior probability by an NN:

$$x_t \rightarrow p(\phi|x_t)$$

- **Justification**:
  – easier learning problem: $\mathcal{O}(10^4)$ labels ϕ vs. vectors $x_t \in \mathbb{R}^{D=40}$
  – well-known result in pattern recognition (but ignored in ASR!)

Hybrid vs. Tandem and Beyond

- **Tandem**:
  - provides high-level, robust and crosslingually generalizing features.
  - known techniques from GMHMM apply (speaker adaptation, discriminative training, etc.)
- **Hybrid**:
  - single model, consistent training.
- **Discussion**:
  - Are they so much different?
  - Relation between Gaussian and log-linear modeling: with pooled covariance only linear features are used: similarity to (unnormalized) softmax layer
  - joint tandem DNN and Gaussian mixture HMM can be viewed as hybrid DNN/HMM: specific topology (combination of linear, sum-/max-pooling and softmax [Tüske & Golik+ 2015])
  - Tandem & Gaussian mixture HMM can be trained jointly [Tüske & Tahir+ 2015] → hybrid
- **Experiments**:
  - [Tüske & Sundermeyer+ 2012, Tüske & Golik+ 2015], ...: similar results for tandem & hybrid
- Towards **deep generative modeling**:
  - combine deep tandem features with deep density models [McDermott 2018]
  - what can be learnt from deep generative modeling in TTS? e.g. [van den Oord+ 2016]
    → utilize for unsupervised training [Tjandra+ 2017]
    → issue: speaker dependence/adaptation [Tjandra+ 2018]

Outline

Discriminative Modeling: from Labels to Frames

- Basically, Bayes decision rule requires modeling of label ([sub]word, ...) **posterior probabilities**
- **Idea**: redefine label sequence on time frame level:

$$p(c_1^N|x_1^T) \leftarrow p(\overline{c}_1^T|x_1^T)$$

with unique mapping from frame-wise to original label sequence $G : \overline{\mathcal{V}}^* \rightarrow \mathcal{V}^*,\ c_1^N = G(\overline{c}_1^T)$

- **Alignment**: marginalize over label boundaries on time frame level

$$\begin{aligned} p(c_1^N|x_1^T) &= \sum_{\overline{c}_1^T} p(\overline{c}_1^T, c_1^N|x_1^T) \\ &= \sum_{\overline{c}_1^T} p(c_1^N|\overline{c}_1^T) p(\overline{c}_1^T|x_1^T) = \sum_{\overline{c}_1^T : G(\overline{c}_1^T) = c_1^N} p(\overline{c}_1^T|x_1^T) \end{aligned}$$

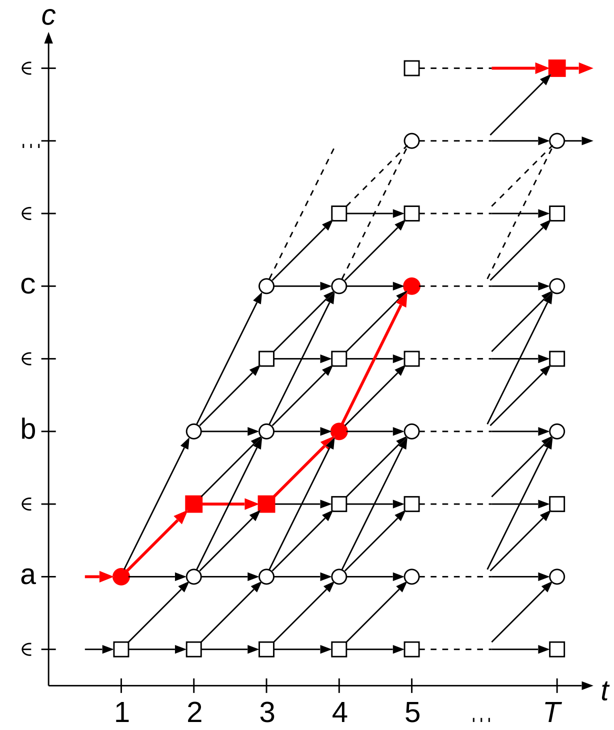
with deterministic frame to label mapping: $p(c_1^N|\overline{c}_1^T) = \begin{cases} 1 & \text{iff} \quad G(\overline{c}_1^T) = c_1^N \\ 0 & \text{otherwise} \end{cases}$

- decompose frame-level posterior $p(\overline{c}_1^T|x_1^T)$ into product over time frames and assume
  - **label independence**: connectionist temporal classification (CTC) [Graves & Fernández$^+$ 2006]
  - **full label context**: RNN-T/recurrent neural aligner [Graves 2012, Sak & Shannon$^+$ 2017]

Connectionist Temporal Classific. (CTC) [Graves & Fernández+ 2006]

- Mapping from frame to label level: extend label set by **blank** symbol "ϵ": $\overline{\mathcal{V}} = \mathcal{V} \cup \{\epsilon\}$
  - blank symbol may be inserted at any point without any effect
  - Adjacent identical labels need to be separated by blank, e.g.:

$$G(\epsilon\mathtt{sspeeee}\epsilon\epsilon\mathtt{e}\epsilon\mathtt{c}\epsilon\mathtt{hhh}\epsilon) = G(\mathtt{spe}\epsilon\mathtt{e}\epsilon\mathtt{cchh}\epsilon) = \mathtt{speech}$$



- Assume **statistical independence** of label sequence

$$p(\overline{c}_1^T | x_1^T) = \prod_{t=1}^{T} p_t(\overline{c}_t | x_1^T)$$

- Related to **hybrid HMM**:
  - two-states per label, 2nd state globally shared for all labels
  - w/o division by state prior
- During training, sum over alignments can be computed with forward-backward algorithm, like the expectation step in the EM algorithm for HMM training.

CTC: Search/Decoding

- w/o language model:
  - leads to independent **frame-by-frame decisions**: trivial
  - with extremely large training set even possible on word level [Soltau & Liao+ 2016]
  - result of statistical independence assumption on label level

$$\operatorname*{argmax}_{\bar{c}_1^T} p(\bar{c}_1^T | x_1^T) = \operatorname*{argmax}_{\bar{c}_1^T} \prod_{t=1}^{T} p_t(\bar{c}_t | x_1^T)$$
$$= \left(\operatorname*{argmax}_{\bar{c}} p_{t=1}(\bar{c} | x_1^T), \ldots, \operatorname*{argmax}_{\bar{c}} p_{t=T}(\bar{c} | x_1^T) \right)$$

  - equivalent to frame-level **word error loss**-based Bayes decision rule [Wessel & Schlüter+ 2001]:

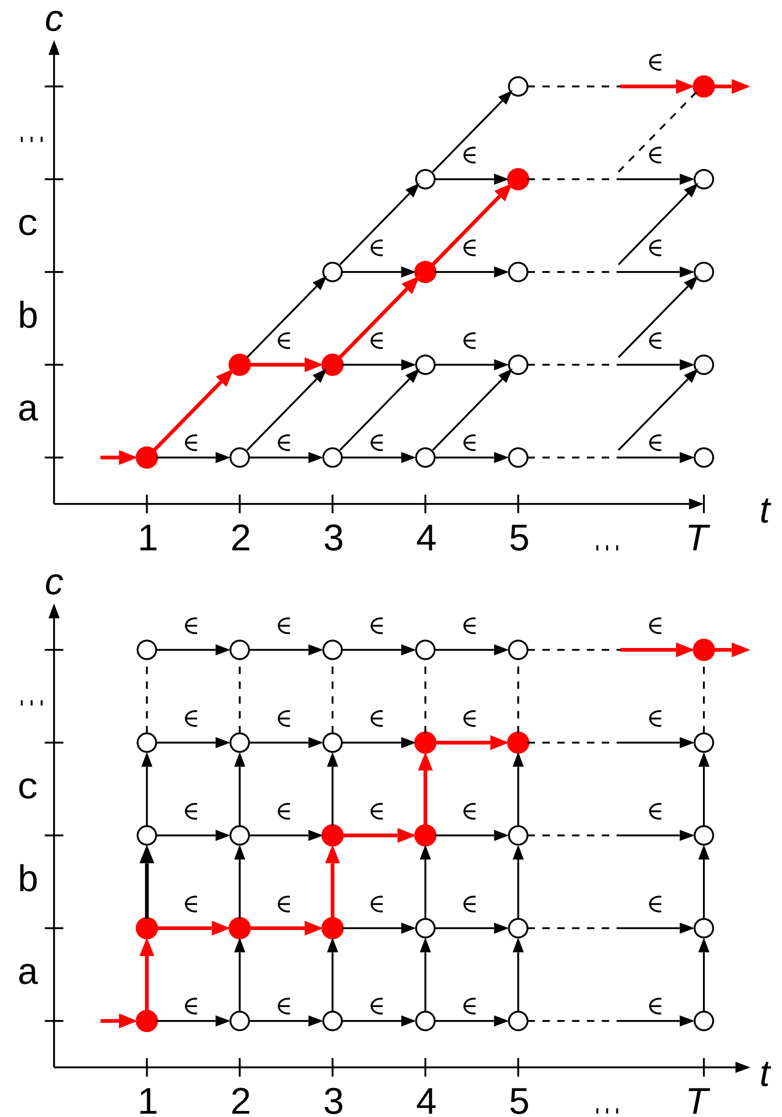
$$\operatorname*{argmin}_{\bar{c}_1^T} \sum_{\hat{c}_1^T} p(\hat{c}_1^T | x_1^T) \cdot \mathcal{C}(\hat{c}_1^T, \bar{c}_1^T) = \operatorname*{argmin}_{\bar{c}_1^T} \sum_{\hat{c}_1^T} p(\hat{c}_1^T | x_1^T) \cdot \sum_{t=1}^{T} (1 - \delta_{\hat{c}_t, \bar{c}_t})$$
$$= \operatorname*{argmax}_{\bar{c}_1^T} \sum_{\tau=1}^{T} p_t(\bar{c}_t | x_1^T)$$
$$= \left(\operatorname*{argmax}_{\bar{c}} p_{t=1}(\bar{c} | x_1^T), \ldots, \operatorname*{argmax}_{\bar{c}} p_{t=T}(\bar{c} | x_1^T) \right)$$

- with language model: CTC used within **hybrid HMM** approach [Miao & Gowayyed+ 2015]
  → standard decoding/search approach (here using WFST)

Recurrent Neural Aligner

- Recurrent neural aligner (RNA) [Sak & Shannon+ 2017]:
  - similar to CTC, but
  - **avoids label independence assumption**:

$$p(\bar{c}_1^T | x_1^T) = \prod_{t=1}^{T} p_t(\bar{c}_t | \bar{c}_1^{t-1}, x_1^T)$$



RNN-Transducer

- RNN-transducer [Graves 2012]:
  - similar to RNA, but does only forward to next frame, if blank label is hypothesized

Search/decoding:

- pursues **tree** of all label sequences
- fixed-size beam pruning

Outline

Directly Hypothesizing Label by Label

- Decompose label sequence posterior probability on a label-by-label level:

$$p(c_1^N|x_1^T) = \prod_{n=1}^{N} p(c_n|c_1^{n-1}, x_1^T)$$

- modeling of **unlimited label context**: can be done by RNN structures (cf. RNN LMs)
- However: how do position-wise label posteriors access/align to corresponding input intervals?
  - encoder/decoder attention and transformer: **attention** in time
  - segmental/inverse HMM: explicit **label boundary** modeling
  - 2D LSTM approach: temporal averaging/**not at all**
- **Advantage**: integrated model, fully exploits interaction between input and label sequence
- **Disadvantage**: training domain integration - domain transfer?

(Non-Latent) Attention-based Encoder/Decoder [Bahdanau & Chorowski+ 2015, Chan & Jaitly+ 2015]

- **decoder input** for n-th label determined by attention process depending on label context:

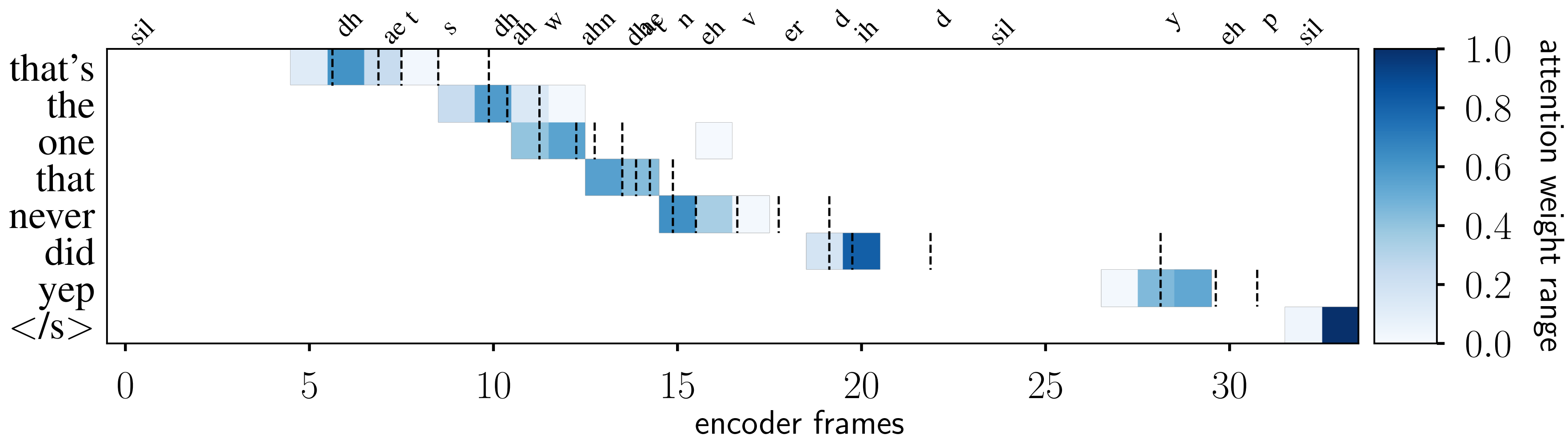
$$p(c_1^N|x_1^T) = \prod_{n=1}^{N} p(c_n|c_1^{n-1}, x_1^T)$$
$$= \prod_{n=1}^{N} p\left(c_n|c_1^{n-1}, \xi(c_1^{n-1}, x_1^T)\right)$$

- **soft attention**: weighted average over encoder output of entire utterance:

$$\xi(c_1^{n-1}, x_1^T) = \sum_{t=1}^{T} \alpha_t(c_1^{n-1}, x_1^T) \cdot x_t$$

- observations/problems:
  - attention determined by context, does **not consider current label**
    $\rightarrow$ attention intervals are not revised after label hypothesization: no recombination
  - left-right asymmetry [Mimura & Sakai+ 2018]
  - competitive performance reported with **sufficiently large training sets**

Attention Visualization



- attention weights: peaky, incomplete coverage of encoder output (depending on downsampling)
  $\Rightarrow$ encoder needs to **temporally compress** information
- informal experiments: attention trained on top of fixed hybrid encoder does not perform
- strong **interaction** of attention and encoder.

- **transformer**: replaces RNN decoder by self-attention, cascades attention [Zhou & Dong^{+} 2018]
  $\rightarrow$ attention issues w.r.t. alignment apply similarly

Segmental/Inverted HMM, Posterior Attention [Lu & Kong+ 2016, Doetsch & Hegselmann+ 2016]

- Idea: label sequence posterior with latent alignment and Markov assumptions:

$$p(c_1^N|x_1^T) = \sum_{t_1^N} p(c_1^N, t_1^N|x_1^T) = \sum_{t_1^N} \prod_{n=1}^{N} p(c_n, t_n|c_1^{n-1}, t_1^{n-1}, x_1^T)$$

$$= \sum_{t_1^N} \prod_{n=1}^{N} p(c_n, t_n|c_1^{n-1}, t_{n-1}, x_1^T) \qquad \text{1}^{\text{st}}\text{-order Markov} \quad \text{joint model (A)}$$

$$= \sum_{t_1^N} \prod_{n=1}^{N} p(c_n|c_1^{n-1}, t_{n-1}, x_1^T) \cdot p(t_n|c_1^{n-1}, c_n, t_{n-1}, x_1^T) \quad \text{target label-dependent (B)}$$

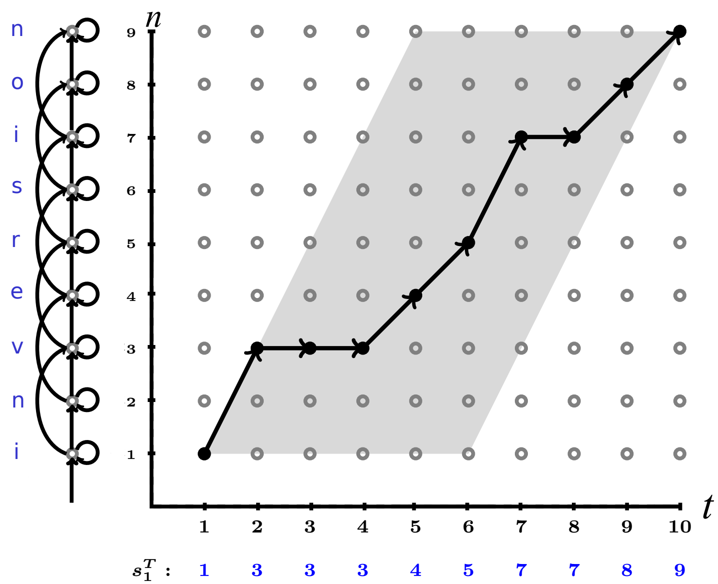
alignment distribution

$$= \sum_{t_1^N} \prod_{n=1}^{N} p(c_n|c_1^{n-1}, t_{n-1}, t_n, x_1^T) \cdot p(t_n|c_1^{n-1}, t_{n-1}, x_1^T) \quad \text{target label-independent}$$

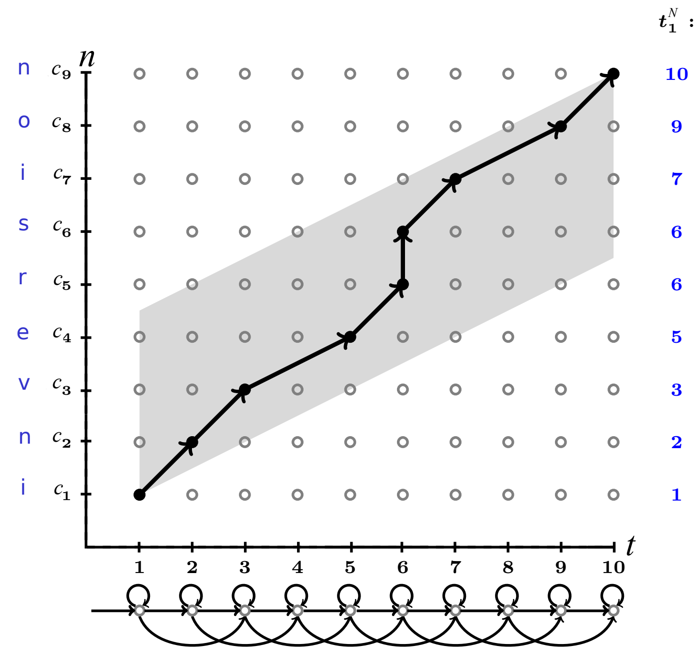
- marginalization of alignment efficiently performed using dynamic programming
- ongoing work: modeling of decoder model distributions for label and alignment

Inverting HMM Alignment

Exemplary toplogies for standard and inverted HMM alignment:



Standard HMM trellis.



Inverted HMM trellis.

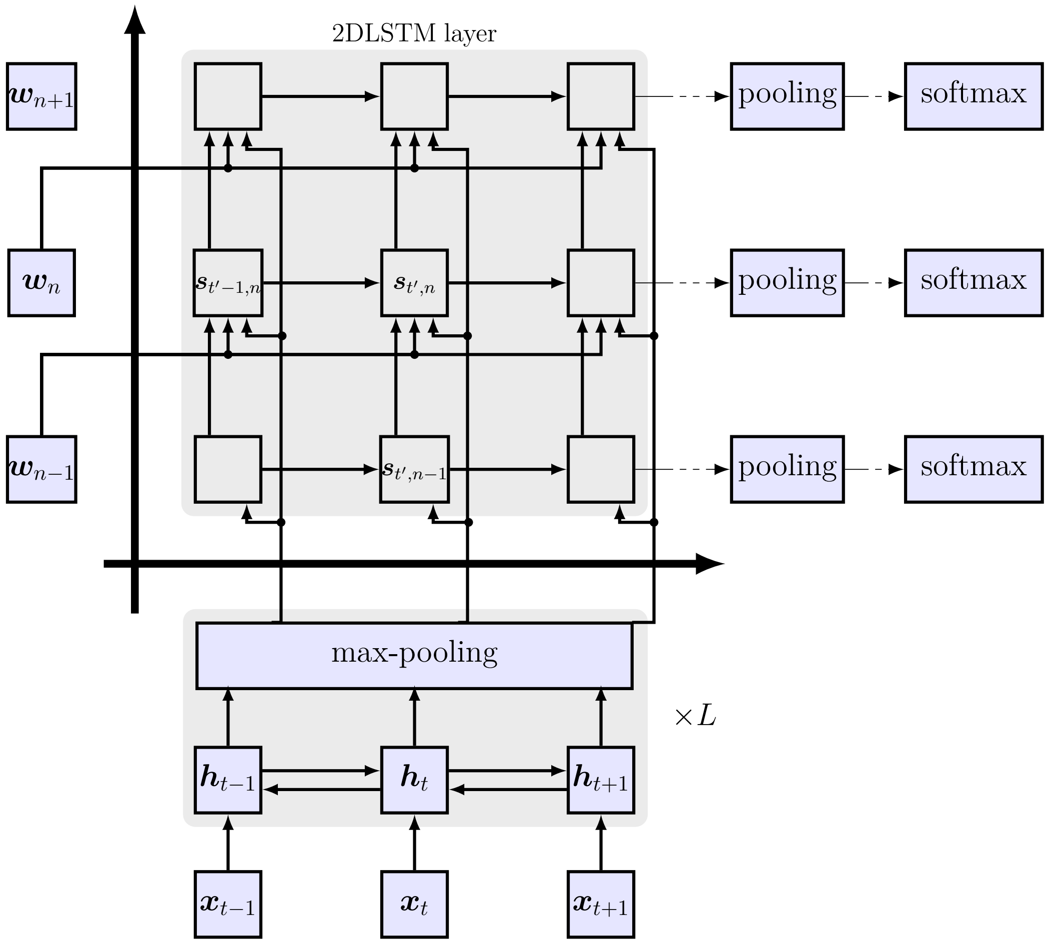
- In MT introduced as **neural HMM** [Wang & Zhu$^+$ 2018]: results similar to attention.
- Introduced as latent generalization of attention:
  posterior attention [Shankar & Sarawagi 2019]: consistently better results reported (BLEU).

2-dim. LSTM [Bahar & Zeyer+ 2019]

- **Idea**: use 2D-LSTM for both label propagation and alignment/ input coverage
- Keep label-synchronous derivation, avoid explicit temporal alignment:

$$p(c_1^N | x_1^T) = \prod_{n=1}^{N} p(c_n | c_1^{n-1}, x_1^T)$$

- **Advantage**:
  exploits 2-dim. structure of input-output relation, completely avoids alignment.
- **Disadvantage**:
  as for soft attention no monotonicity or locality constraints.



2D-LSTM architecture avoiding attention.

Outline

Current Results for new Architectures

- Training: LibriSpeech 1000h, Switchboard 300h
- CDP: context-dependent phonemes (generalized triphone states)
- BPE: subwords based on byte-pair encoding, 1000 merges

acoustic model		language model		WER [%]	
				Switchboard	LibriSpeech
approach	labels	labels	approach	Hub5 '00	test-other
inv. HMM	CDP	words	4-gram	13.0	-
2D-LSTM	BPE		none	10.6	-
attention				9.9	10.3
			LSTM	9.3	8.2
			Transformer	9.2	7.5
hybrid	CDP	words	4-gram	8.1	8.8
			LSTM	6.7	5.5
			Transformer	**6.6**	**5.0**

(Librispeech results, hybrid: [Lüscher & Beck+ 2019], attention: unpublished 2019)
(Switchboard results, hybrid: [Kitza & Schlüter+ 2019], attention: unpublished 2019)
(Inv. HMM results: [Beck & Hannemann+ 2018], work in progress)
(2D-LSTM results: work in progress following [Bahar & Zeyer+ 2019])

Performance as a Function of Training Data Amount

GMM/HMM vs. hybrid BLSTM/HMM vs. BLSTM/attention:
Comparison on LibriSpeech, dev-clean

amount training data [h]	WER [%] dev-clean		
	HMM		Attention
	GMM AM + 4g LM	BLSTM AM + LSTM LM	
10	13.0	9.2	>100
50			23.0
100	9.7	5.1	10.1
1000*	7.6	2.2	2.9

(* Resuls for 1000h from [Lüscher & Beck+ 2019])

Results on LibriSpeech Test

Results published at this Interspeech* and coming ASRU^o 2019

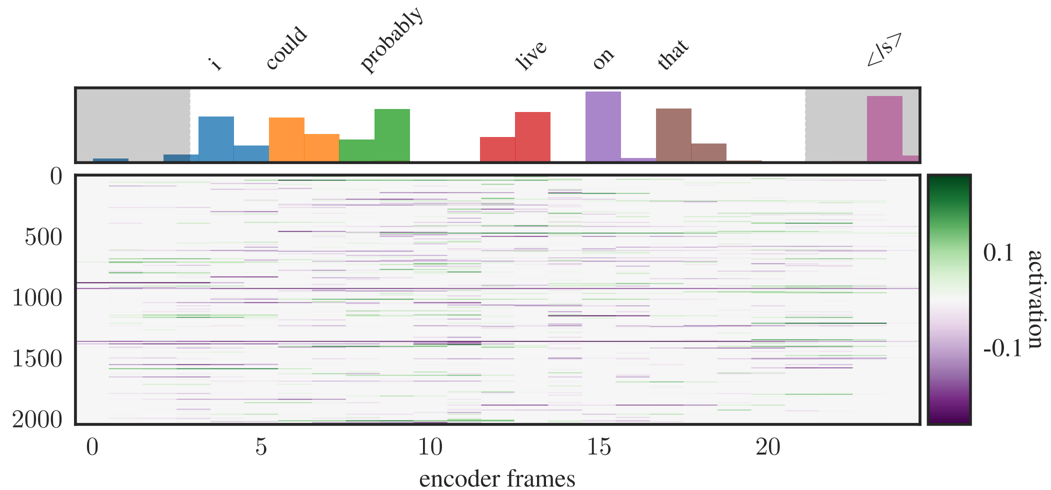
data augm.	approach	encoder	LM	WER [%] clean	other	publication
no	CTC	CNN	6gram	3.3	9.6	[Li & Lavrukhin+ 2019]*
	attention	TDS conv	CNN	3.3	9.8	[Hannun & Lee+ 2019]*
yes	CTC	CNN	Trafo	2.8	7.8	[Li & Lavrukhin+ 2019]*
	attention	LSTM	LSTM	2.5	8.0	[Tüske & Audhkhasi+ 2019]*
				2.5	5.8	[Park & Chan+ 2019]*
		Trafo	Trafo	2.8	7.4	[Zeyer & Bahar+ 2019]^o
		LSTM		2.4	8.2	[Kim & Shin+ 2019]*
	CTC+att'n	Trafo	RNN	2.6	5.7	[Karita & Chen+ 2019]^o
no	hybrid	LSTM	LSTM	2.6	5.5	[Lüscher & Beck+ 2019]*
			Trafo	**2.3**	**5.0**	

Encoder

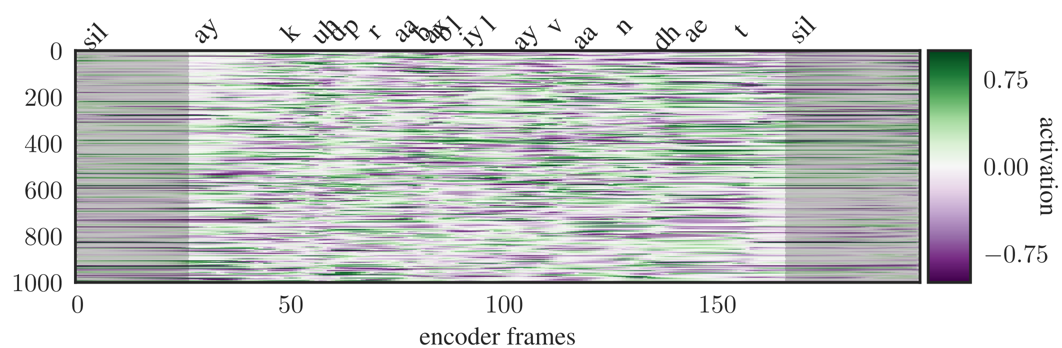
- even hybrid model can be seen as:
  - encoder (up to last hidden layer)
  - decoder (output activation+softmax: log-linear layer)

- formally, encoder modeling similar for all cases, e.g.using **deep bidirectional LSTMs**, some variation:
  - temporal sub-sampling
  - layer sizes

- however, parameterization after training may vary strongly, e.g. attention vs. hybrid:

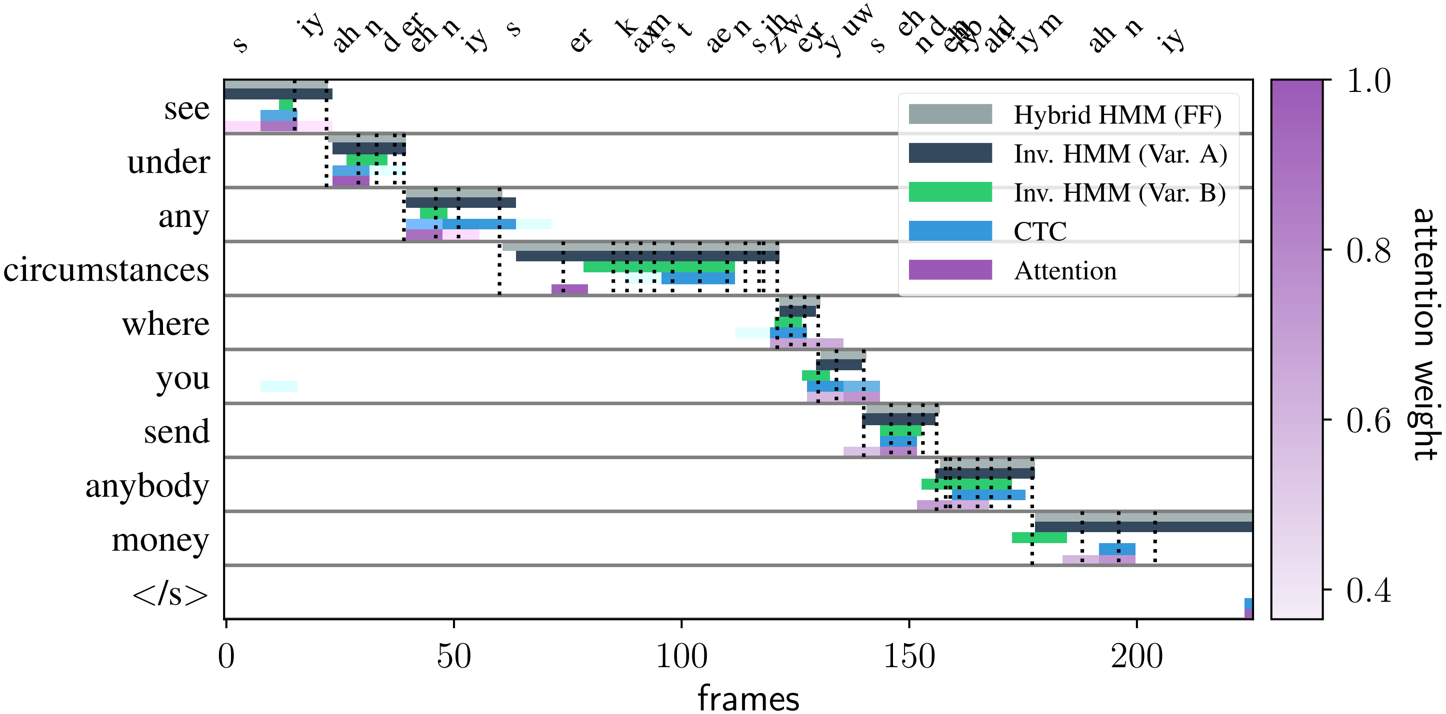


Attention and corresponding encoder.



Phoneme positions and hybrid DNN/HMM encoder.

Alignment



Var. A: first-order joint model

Var. B: separate length model

Comparison of alignment/attention for exemplary SWB utterance.

- attention strongly **localized**, variation in label length covered by attention positioning
  $\rightarrow$ alignment: **interaction** between attention and encoder!
  $\rightarrow$ encoding: necessarily differs between hybrid and inverted HMM
- depending on modeling, inverted HMM aligns similar to hybrid HMM

Vocabulary Modeling

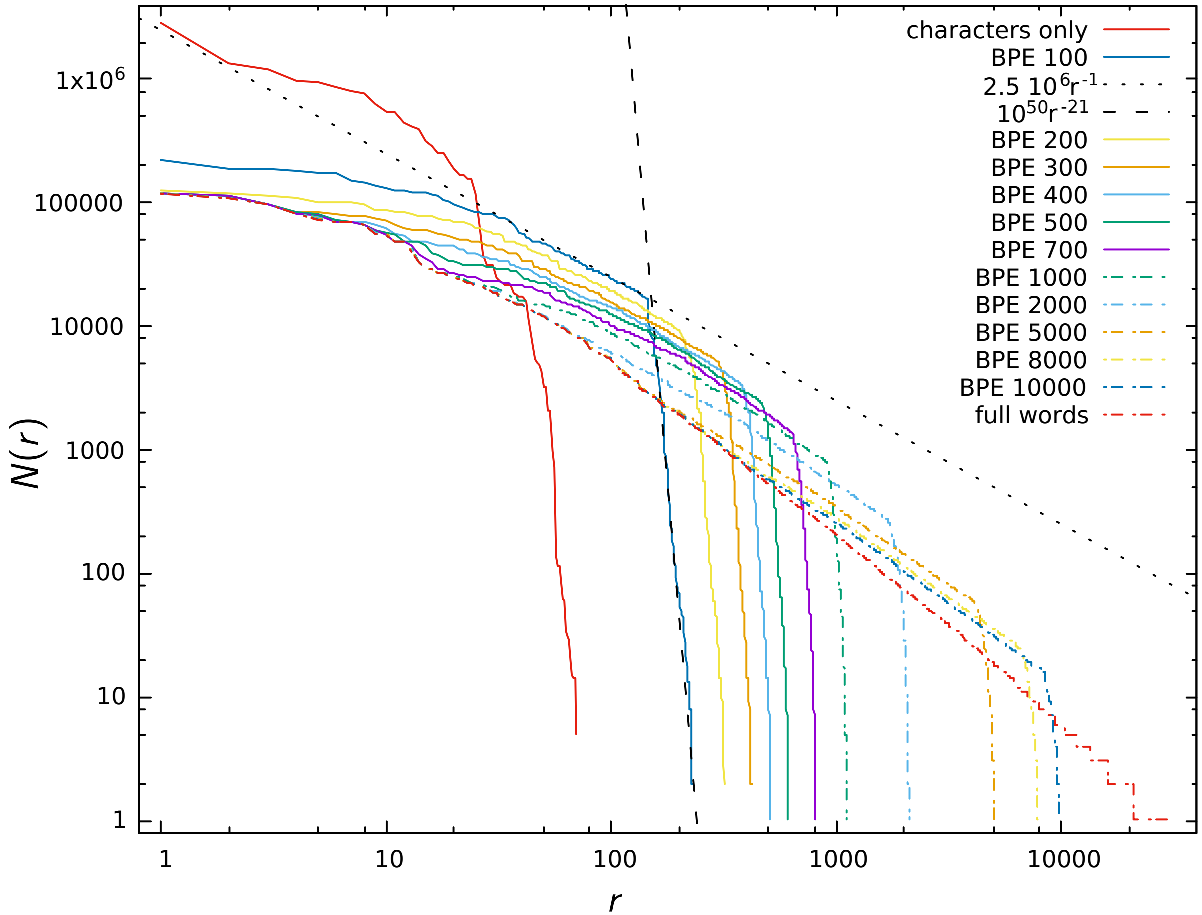
Goal:

- discard intermediate modeling based on pronunications
  → **avoid pronunciation lexicon**
- enable direct vocabulary modeling
  → how to cover **words unseen** during training?
  e.g. character-based, even for HMM, cf. e.g. [Kanthak & Ney 2002], or Babel project

Approach:

- decompose words into subwords
  → enables **open vocabulary**, provided all characters are included
  (explicitly or implicitly over subsequences)
- byte-pair encoding (BPE) [Sennrich, Haddow+ 2016]
  - originally data compression approach
  - successive agglomeration of frequent character (byte) pairs
  - short BPE units: good statistics, but acoustic realization (pronunciation) possibly ambiguous
  - long BPE units/full words: proper pronunciation, but much longer tail of infrequent units

Beyond Zipf's Law: Byte Pair Encoding

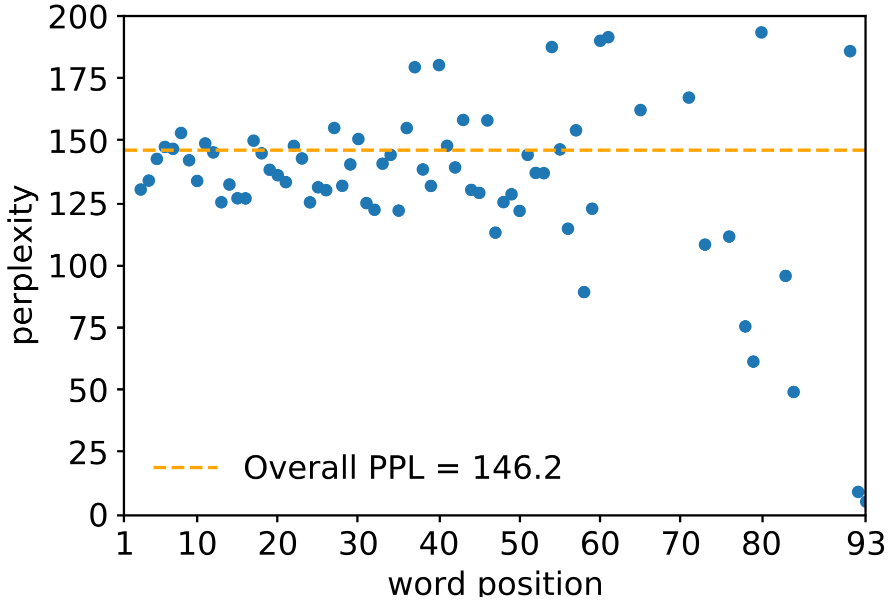


Label rank r vs. frequency $N(r)$ for different vocabularies (Switchboard task).

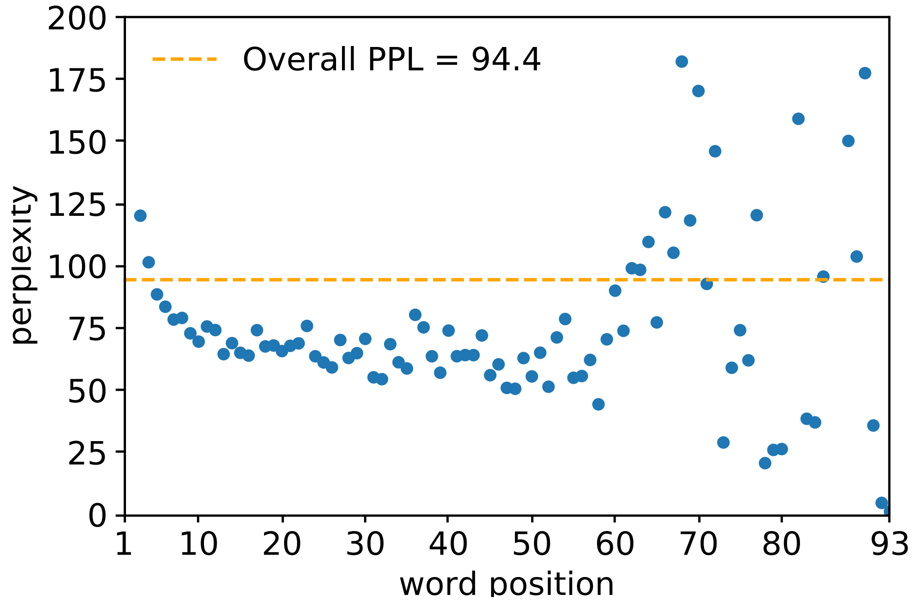
(Dichotomy in Zipf's Law: cf. [Montemurro 2001].)

Label Positional Perplexity Trend

- LibriSpeech Dev clean+other perplexities per word position:



Word 4-gram LM: approx. stationary.



Word LSTM LM: clear initial trend.

- Full/recurrent word context models show **trend** over word positions.
- Supports "middle-out decoding" approach proposed at this NeurIPS in [Mehri & Sigal 2018]
- Might partly explain directional asymmetry considered in [Mimura & Sakai^{+} 2018].

Search/Decoding, Domain-Dependence

- Various HMM approaches and CTC with LM: search includes **alignment optimization**.
- Standard beam search with **relative pruning**, look-ahead methods and dynamic search spaces.
- Attention: search only on label level, attention is not globally optimized: locally determined by the label history: constitutes **intermediate decisions** to some extent.
- Label-synchronous decoding: how to perform pruning? Which hypotheses are comparable? Relation to input coverage?
- Size of search space varies with model quality and with input properties, search in end-to-end systems often is reduced to small fixed-size beams.

Separate audio and text data resources:

- clear separation in **standard decomposition** into acoustic/language model
- speech chain allows inclusion of separate textual data during training [Tjandra$^+$ 2017]: interpret concatenation of TTS and ASR (and vice versa) as text (speech audio) autoencoders
- [Sriram & Jun$^+$ 2018] includes **LM in decoder training** to prevent that the decoder implicitly learns LM information

Outline

Conclusions

Common characterization of end-to-end systems:

- directly convert input (audio signal) into output (word sequences)
- do not involve **intermediate representations** (ASR: phoneme set, pronunication lexicon)
- can be trained from scratch end-to-end to optimize performance measure (ASR: word error rate)

Discussion:

- **Integrated decision** end-to-end based on all knowledge sources:
  natural goal of **statistical approach** to ASR - caveats: beam search, search complexity?
- Existing **knowledge sources** (e.g. signal processing, phonetic, temporal segmentation, existing models like multilingual features, etc.) may be viewed as additional (possibly noisy or mismatched) "data" - using it may still help, especially if primary training data is sparse.
- Internal structuring provides **intermediate representations** that enable internal model analysis to some extent.
- taking **training from scratch** literally would also exclude pretraining or any hyperparameter optimization (aka repeated training and testing on held out data).
- Training hierarchically with **intermediate representations** and corresponding objectives provides potential modes of initialization, regularization, and analysis.
- Transition between training from scratch and using **prior knowledge** needed: supported by machine learning methods.

Current Situation

Training

- Any ASR system today is sequence discriminative trainable.
- However: pretraining/prior training with different objective might be necessary.
- Hyperparameter optimization concerns all approaches.
- Varying amounts of training data:
  - Insertion of external knowledge sources?
  - Transition from standard to novel end-to-end models?

Recognition:

- Strictly speaking, only CTC *fully* searchable (but...).
- Small vocabulary and short context LM: no pruning needed.
- All others not strictly optimal, incl. end-to-end:
  - Beam search, pruning: global optimum not guaranteed.
  - Exponential search tree with RNN LM and/or decoder.
  - How does an end-to-end system indicate uncertainty?
    $\rightarrow$ Calibration [Guo & Pleiss$^+$ 2017] needed?
  - "Two-Pass End-to-End" Speech Recognition [Sainath & Pang$^+$ 2019]

Thank you for your attention!

Outline

References

- S. Shankar, S. Sarawagi: "Posterior Attention Models for Sequence to Sequence Learning," *International Conference on Learning Representations*, New Orleans, LA, May 2019.
- P. Bahar, A. Zeyer, R. Schlüter, H. Ney: "On using 2D Sequence-to-Sequence Models for Speech Recognition," *IEEE Intern. Conf. on Acoustics, Speech, and Signal Processing*, Brighton, UK, May 2019.
- D. Bahdanau, J. Chorowski, D. Serdyuk, P. Brakel, Y. Bengio: "End-to-End Attention-based Large Vocabulary Speech Recognition," *arXiv preprint*, arXiv:1508.04395, Aug. 2015.
- L. R. Bahl, F. Jelinek, R. L. Mercer: A Maximum Likelihood Approach to Continuous Speech Recognition. IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. 5, pp. 179-190, March 1983.

References

- L. Bahl, S. De Gennaro, P. Gopalakrishnan, R. Mercer: “A Fast Approximate Acoustic Match for Large Vocabulary Speech Recognition,” *IEEE Transactions on Speech and Audio Processing*, Vol. 1, No. 1, pp. 59–67, Jan. 1993.
- T. Bayes: “An Essay towards solving a problem in the doctrine of chances,” *Philosophical Transactions of the Royal Society of London*, Vol. 53, pp. 370–418, 1763, (Reprinted in *Biometrika*, Vol. 45, No. 3/4, pp. 293–315, Dec. 1958).
- E. Beck, M. Hannemann, P. Doetsch, R. Schlüter, H. Ney: “Segmental Encoder-Decoder Models for Large Vocabulary Automatic Speech Recognition,” *Interspeech*, Hyderabad, India, Sep. 2018.
- Y. Bengio, R. Ducharme, P. Vincent: A neural probabilistic language model. Advances in Neural Information Processing Systems (NIPS), pp. 933-938, Denver, CO, Nov. 2000.
- H. Bourlard, C. J. Wellekens: 'Links between Markov Models and Multilayer Perceptrons', in D.S. Touretzky (ed.): ”Advances in Neural Information

References

Processing Systems I", Morgan Kaufmann Pub., San Mateo, CA, pp.502-507, 1989.

- H. Bourlard, N. Morgan: *Connectionist Speech Recognition: a Hybrid Approach*, Kluwer Academic Publishers, Norwell, MA, 1993.
- J. S. Bridle: Probabilistic Interpretation of Feedforward Classification Network Outputs with Relationships to Statistical Pattern Recognition, in F. Fogelman-Soulie, J. Herault (eds.): 'Neuro-computing: Algorithms, Architectures and Applications', NATO ASI Series in Systems and Computer Science, Springer, New York, 1989.
- M.A. Castano, E. Vidal, F. Casacuberta: Inference of stochastic regular languages through simple recurrent networks. IEE Colloquium on Grammatical Inference: Theory, Applications and Alternatives, pp. 16/1-6, Colchester, UK, April 1993.
- W. Chan, N. Jaitly, Q. V. Le, O. Vinyals: "Listen, Attend and Spell," *arXiv preprint*, arXiv:1508.01211, Aug. 2015.

References

- G. E. Dahl, D. Yu, L. Deng, A. Acero: Context-Dependent Pre-Trained Deep Neural Networks for Large-Vocabulary Speech Recognition. IEEE Tran. on Audio, Speech and Language Processing, Vol. 20, No. 1, pp. 30-42, Jan. 2012.
- S. Davis, P. Mermelstein: “Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol. ASSP-28, No. 4, pp. 357–366, Aug. 1980.
- P. Doetsch, S. Hegselmann, R. Schlüter, and H. Ney: “Inverted HMM - a Proof of Concept,” *Neural Information Processing Systems (NIPS) Workshop*, Barcelona, Spain, Dec. 2016.
- V. Fontaine, C. Ris, J.-M. Boite: Nonlinear discriminant analysis for improved speech recognition, Eurospeech, Rhodes, Greece, Sept. 1997.

References

- J. Fritsch, M. Finke, A. Waibel: Adaptively Growing Hierarchical Mixtures of Experts. NIPS, Advances in Neural Information Processing Systems 9, MIT Press, pp. 459-465, 1997.
- F. A. Gers, N. N. Schraudolph, J. Schmidhuber: Learning precise timing with LSTM recurrent networks. Journal of Machine Learning Research, Vol. 3, pp. 115-143, 2002.
- V. Goel, W. Byrne: "Minimum Bayes Risk Automatic Speech Recognition," *Computer Speech and Language*, Vol. 14, No. 2, pp. 115–135, April 2000.
- P. Golik, Z. Tüske, R. Schlüter, H. Ney: "Convolutional Neural Networks for Acoustic Modeling of Raw Time Signal in LVCSR," *Interspeech*, pp. 26-30, Dresden, Germany, September 2015.
- P. Golik, Z. Tüske, R. Schlüter, H. Ney: "Multilingual Features Based Keyword Search for Very Low-Resource Languages," *Interspeech*, pp. 1260–1264, Dresden, Germany, Sep. 2015.

References

- A. Graves, S. Fernández, F. Gomez, J. Schmidhuber: "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," *Int. Conf. on Machine Learning (ICML)*, pp. 369–376, Helsinki, Finland, June 2006.
- A. Graves, M. Liwicki, H. Bunke, J. Schmidhuber, S. Fernandez: "Unconstrained online handwriting recognition with recurrent neural networks," In Advances in Neural Information Processing Systems, Vol. 20. MIT Press, 2008.
- A. Graves: "Sequence Transduction with Recurrent Neural Networks," *arXiv preprint*, arXiv:1211.3711, Nov. 2012.
- F. Grézl, M. Karafiát, S. Kontár, J. Cernocký: "Probabilistic and Bottle-neck Features for LVCSR of Meetings," *IEEE Intern. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 757–760, Honolulu, HI, April 2007.
- C. Guo, G. Pleiss, Y. Sun, K.Q. Weinberger: "On Calibration of Modern Neural Networks," *arXiv preprint*, arXiv:1706.04599, Jun. 2017.

References

- K. J. Han, A. Chandrashekaran, J. Kim, I. Lane: "The CAPIO 2017 Conversational Speech Recognition System," *arXiv preprint*, arXiv:1801.00059, Jan. 2018.
- A. Hannun, A. Lee, Q. Xu, R. Collobert: "Sequence-to-Sequence Speech Recognition with Time-Depth Separable Convolutions," *Interspeech*, Graz. Austria, Sep. 2019.
- H. Hermansky, D. Ellis, S. Sharma: "Tandem Connectionist Feature Extraction for Conven- tional HMM Systems," *IEEE Intern. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Vol. 3, pp. 1635–1638, Istanbul, Turkey, June 2000.
- G. Hinton, S. Osindero, Y. Teh: "A Fast Learning Algorithm for Deep Belief Nets," *Neural Computation*, Vol. 18, No. 7, pp. 1527Â-1554, July 2006.
- S. Hochreiter, J. Schmidhuber: "Long short-term memory," *Neural Computation*, Vol. 9, No. 8, pp. 1735–1780, Nov. 1997.

References

- T. Hori, Y. Kubo, A. Nakamura: "Real-time one-pass decoding with recurrent neural network language model for speech recognition," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6364–6368, Florence, Italy, May 2014.
- K. Irie, A. Zeyer, R. Schlüter, H. Ney: "Language Modeling with Deep Transformer Architectures" *Interspeech*, Graz. Austria, Sep. 2019.
- F. Jelinek, L. R. Bahl, R. L. Mercer: "Design of a Linguistic Statistical Decoder for the Recognition of Continuous Speech," *IEEE Transactions on Information Theory*, Vol. IT-21, No. 3, May 1975.
- N. Kanda, C. Boeddeker, J. Heitkaemper, Y. Fujita, S. Horiguchi, K. Nagamatsu, R. Häb-Umbach: "Guided Source Separation Meets a Strong ASR Backend: Hitachi/Paderborn University Joint Investigation for Dinner Party ASR," *Interspeech*, Graz. Austria, Sep. 2019.

References

- S. Kanthak, H. Ney: "Context-Dependent Acoustic Modeling Using Graphemes for Large Vocabulary Speech Recognition," *IEEE Intern. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 845–848, Orlando, FL, May 2002.
- S. Karita, N. Chen, T. Hayashi, T. Hori, H. Inaguma, Z. Jiang, M. Someki, N. E. Yalta Soplin, R. Yamamoto, X. Wang, S. Watanabe, T. Yoshimura, W. Zhang: "A Comparative Study on Transformer vs RNN in Speech Applications," IEEE Automatic Spreech Recongnition and Understanding Workshop(ASRU), Singapore, Dec. 2019, *to appear*.
- C. Kim, M. Shin, A. Garg, D. Gowda: "Improved vocal tract length perturbation for a state-of-the-art end-to-end speech recognition system," *Interspeech*, Graz. Austria, Sep. 2019.
- M. Kitza, R. Schlüter, H. Ney: "Cumulative Adaptation of BLSTM Acoustic Models using I-Vectors and Cluster Dependent Transformations," *Interspeech*, Graz. Austria, Sep. 2019.

References

- R. Kneser, H. Ney: "Improved clustering techniques for class-based statistical language modelling," *Eurospeech*, Vol. 93, pp. 973–976, Berlin, Germany, Sep. 1993.
- Y. LeCun, Y. Bengio: Word-level training of a handwritten word recognizer based on convolutional neural networks. Int. Conf. on Pattern Recognition, Jerusalem, Israel, pp. 88-92, Oct. 1994.
- J. Li, V. Lavrukhin, B. Ginsburg, R. Leary, O. Kuchaiev, J. M. Cohen, H. Nguyen, R. Teja Gadde: "Jasper: An End-to-End Convolutional Neural Acoustic Model," *Interspeech*, Graz. Austria, Sep. 2019.
- P. J. Liu, M. Saleh, E. Pot, B. Goodrich, R. Sepassi, L. Kaiser, N. Shazeer: "Generating wikipedia by summarizing long sequences," Int. Conf. on Learning Representations (ICLR), Vancouver, Canada, April 2018.

References

- L. Lu, L. Kong, C. Dyer, N. A. Smith, S. Renals: "Segmental Recurrent Neural Networks for End-to-End Speech Recognition," *Interspeech*, pp. 385–389, Sep. 2016.
- C. Lüscher, E. Beck, K. Irie, M. Kitza, W. Michel, A. Zeyer, R. Schlüter, H. Ney: "RWTH ASR Systems for Librispeech: Hybrid vs. Attention," *Interspeech*, Graz. Austria, Sep. 2019.
- E. McDermott: "A Deep Generative Acoustic Model for Compositional Automatic Speech Recognition," *Neural Information Processing Systems (NeurIPS) Workshop: Interpretability and Robustness in Audio, Speech, and Language*, Montreal, QC, Canada, Dec. 2018.
- S. Mehri, L. Sigal: "Middle-Out Decoding," *Conference on Neural Information Processing Systems (NeurIPS)*, Montreal, QC, Canada, Dec. 2018.
- Z. Meng, Y. Gaur, J. Li, Y. Gong: "Speaker Adaptation for Attention-Based End-to-End Speech Recognition," *Interspeech*, Graz. Austria, Sep. 2019.

References

- E. Tsunoo, Y. Kashiwagi, S. Asakawa, T. Kumakura: "End-to-end Adaptation with Backpropagation through WFST for On-device Speech Recognition System," *Interspeech*, Graz. Austria, Sep. 2019.
- T. Menne, I. Sklyar, R. Schlüter, H. Ney: "Speaker Separation with Deep Clustering as Preprocessing For Automatic Speech Recognition of Sparsely Overlapping Speech" *Interspeech*, Graz. Austria, Sep. 2019.
- A. Merboldt, A. Zeyer, R. Schlüter, H. Ney: "An Analysis of Local Monotonic Attention Variants" *Interspeech*, Graz. Austria, Sep. 2019.
- Y. Miao, M. Gowayyed, F. Metze: "EESEN: End-to-End Speech Recognition using Deep RNN Models and WFST-based Decoding," *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pp. 167–174, Scottsdale, AZ, Dec. 2015.

References

- W. Michel, R. Schlüter, H. Ney: "Comparison of Lattice-Free and Lattice-Based Sequence Discriminative Training Criteria for LVCSR" *Interspeech*, Graz. Austria, Sep. 2019.
- T. Mikolov, M. Karafiat, L. Burget, J. ernocky, S. Khudanpur: "Recurrent neural network based language model," *Interspeech*, pp. 1045–1048, Makuhari, Chiba, Japan, Sep. 2010.
- M. Mimura, S. Sakai, T. Kawahara: "Forward-Backward Attention Decoder," *Interspeech*, Hyderabad, India, Sep. 2018.
- M. Mohri, M. Riley: "Integrated Context Dependent Networks in Very Large Vocabulary Speech Recognition," *Europ. Conf. on Speech Communication (EuroSpeech)*, pp. 811–814, Budapest, Hungary, Sept. 1999.
- A. Montemurro: "Beyond the Zipf-Mandelbrot law in quantitative linguistics," *Physica A*, Vol. 300, pp. 567–578, 2001.

References

- M. Nakamura, K. Shikano: "A Study of English Word Category Prediction Based on Neural Networks," *IEEE Intern. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 731–734, Glasgow, UK, May 1989.
- H. Ney, R. Haeb-Umbach, B.-H. Tran, M. Oerder: "Improvements in beam search for 10000-word continuous speech recognition," *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol. 1, pp. 9–12, San Francisco, CA, 1992.
- S. Ortmanns, H. Ney, N. Coenen, "Language Model Look-Ahead for Large Vocabulary Speech Recognition," *Intern. Conf. on Spoken Language Processing (ICSLP)*, pp. 2095–2098, Philadelphia, PA, Oct. 1996.
- D. Palaz, R. Collobert, M. Magimai-Doss: "Estimating phoneme class conditional probabilities from raw speech signal using convolutional neural networks," *Interspeech*, pp. 1766–1770, Lyon, France, Aug. 2013.

References

- D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, Q. V. Le: “SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition,” *Interspeech*, Graz. Austria, Sep. 2019.
- T. Raissi, E. Beck, R. Schlüter, H. Ney: “Context-Dependent Acoustic Modelling Without Classification and Regressions Trees,” *Workshop for Young Female Researchers in Speech Science & Technology, Interspeech*, Graz. Austria, Sep. 2019.
- A. J. Robinson: An Application of Recurrent Nets to Phone Probability Estimation. IEEE Trans. on Neural Networks, Vol. 5, No. 2, pp. 298-305, March 1994.
- D. Rybach, H. Ney, R. Schlüter: “Lexical Prefix Tree and WFST: A Comparison of Two Dynamic Search Concepts for LVCSR,” *IEEE Trans. on Audio, Speech, and Language Processing*, Vol. 21, No. 6, pp. 1295–1307, June 2013.

References

- T. N. Sainath, R. Pang, D. Rybach, Y. He, R. Prabhavalkar, W. Li, M. Visontai, Q. Liang, T. Strohman, Y. Wu, I. McGraw, C.-C. Chiu: "Two-Pass End-to-End Speech Recognition," *Interspeech*, Graz. Austria, Sep. 2019.
- T.N. Sainath, , R.J. Weiss, K.W. Wilson, A. Narayanan, M. Bacchiani: "Speaker Location and Microphone Spacing Invariant Acoustic Modeling from Raw Multichannel Waveforms," *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pp. 30–36, Scottsdale, AZ, Dec. 2015.
- H. Sak, M. Shannon, K. Rao, F. Beaufays: "Recurrent Neural Aligner: An Encoder-Decoder Neural Network Model for Sequence to Sequence Mapping," *Interspeech*, pp. 1298–1302, Stockholm, Sweden, Aug. 2017.
- G. Saon, G. Kurata, T. Sercu, K. Audhkhasi, S. Thomas, D. Dimitriadis, X. Cui, B. Ramabhadran, M. Picheny, L.L. Lim, B. Roomi, P. Hall: "English Conversational Telephone Speech Recognition by Humans and Machines," *arXiv preprint*, arxiv:1703.02136, Mar. 2017.

References

- R. Schlüter, M. Nussbaum-Thom, H. Ney: Does the Cost Function Matter in Bayes Decision Rule? IEEE Trans. PAMI, No. 2, pp. 292–301, Feb. 2012.
- H. Schwenk: Continuous space language models. Computer Speech and Language, Vol. 21, No. 3, pp. 492–518, July 2007.
- F. Seide, G. Li, D. Yu: “Conversational Speech Transcription using Context-Dependent Deep Neural Networks,” *Interspeech*, pp. 437–440, Florence, Italy, Aug. 2011.
- R. Sennrich, B. Haddow, A. Birch: “Neural machine translation of rare words with subword units,” in *ACL*, pp. 1715–1725, Berlin, Aug. 2016.
- H. Soltau, H. Liao, H. Sak: “Neural Speech Recognizer: Acoustic-to-Word LSTM Model for Large Vocabulary Speech Recognition,” *arXiv preprint*, arxiv:1610.09975, Oct. 2016.

References

- A. Sriram, H. Jun, S. Satheesh, A. Coates: “Cold Fusion: Training Seq2Seq Models Together with Language Models,” *Interspeech*, Hyderabad, India, Sep. 2018.
- V. Steinbiss, B.-H. Tran, H. Ney: “Improvements in Beam Search,” *Int. Conf. on Spoken Language Processing (ICSLP)*, pp. 2143–2146, Yokohama, Japan, Sept. 1994.
- A. Stolcke, F. Grézl, M.-Y. Hwang, X. Lei, N. Morgan, D. Vergyri: “Cross-domain and cross-language portability of acoustic features estimated by multilayer perceptrons,” 'textitIEEE Int. Conf. on Acoustics, Speech and Signal Processing, pp. 321–324, Toulouse, France, May 2006.
- M. Sundermeyer, R. Schlüter, H. Ney: “LSTM neural networks for language modeling,” Interspeech, pp. 194–197, Portland, OR, Sep. 2012.

References

- M. Sundermeyer, Z. Tüske, R. Schlüter, H. Ney: "Lattice Decoding and Rescoring with Long-Span Neural Network Language Models," *Interspeech*, pp. 661–665, Singapore, Sep. 2014.
- A. Tjandra, S. Sakti, S. Nakamura: "Listening while Speaking: Speech Chain by Deep Learning," *IEEE Automatic Speech Recognition and Understanding (ASRU) Workshop*, Okinawa, Japan, Dec. 2017.
- A. Tjandra, S. Sakti, S. Nakamura: "Machine Speech Chain with One-shot Speaker Adaptation," *Interspeech*, Hyderabad, India, Sep. 2018.
- Z. Tüske, K. Audhkhasi, G. Saon: "Advancing Sequence-to-Sequence based Speech Recognition," *Interspeech*, Graz. Austria, Sep. 2019.
- Z. Tüske, P. Golik, R. Schlüter, H. Ney: "Acoustic Modeling with Deep Neural Networks Using Raw Time Signal for LVCSR," *Interspeech*, pp. 890–894, Singapore, September 2014.

References

- Z. Tüske, J. Pinto, D. Willett, R. Schlüter: "Investigation on Cross- and Multilingual MLP Features under Matched and Mismatched Acoustical Conditions," *IEEE Intern. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 7349–7353, Vancouver, BC, Canada, May 2013.
- Z. Tüske, P. Golik, R. Schlüter, H. Ney: "Speaker Adaptive Joint Training of Gaussian Mixture Models and Bottleneck Features," *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 596–603, Scottsdale, AZ, Dec. 2015.
- Z. Tüske, M. Sundermeyer, R. Schlüter, H. Ney: "Context-Dependent MLPs for LVCSR: TANDEM, Hybrid or Both?" *Interspeech*, pp. 18–21, Portland, OR, Sept. 2012.
- Z. Tüske, M. A. Tahir, R. Schlüter, H. Ney: "Integrating Gaussian Mixtures into Deep Neural Networks: Softmax Layer with Hidden Variables," *IEEE Intern.*

Conf. on Acoustics, Speech, and Signal Processing (ICASSP), pp. 4285–4289, Brisbane, Australia, Apr. 2015.

- Z. Tüske, K. Irie, R. Schlüter, H. Ney: "Investigation on Log-Linear Interpolation of Multi-Domain Neural Network Language Model," *IEEE Inter. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 6005–6009, Shanghai, China, Mar. 2016.
- P. E. Utgoff, D. J. Stracuzzi: "Many-layered learning," Neural Computation, Vol. 14, No. 10, pp. 2497-2539, Oct. 2002.
- A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, K. Kavukcuoglu: "WaveNet: A Generative Model for Raw Audio," *arXiv preprint*, arxiv:1609.03499, Sep. 2016.
- A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, K. L. Lang: Phoneme Recognition: Neural Networks vs. Hidden Markov Models. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), New York, NY, pp.107-110, April 1988.

References

- W. Wang, D. Zhu, T. Alkhouli, Z. Gan, H. Ney: "Neural Hidden Markov Model for Machine Translation," *Annual Meeting of the Assoc. for Computational Linguistics (ACL)*, Melbourne, Australia, July 2018.
- F. Wessel, R. Schlüter, H. Ney: "Explicit Word Error Minimization using Word Hypothesis Posterior Probabilities," Proc. *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 33–36, Salt Lake City, Utah, May 2001.
- W. Xiong, J. Droppo, X. Huang, F. Seide, M. Seltzer, A. Stolcke, D. Yu, G. Zweig: Achieving Human Parity in Conversational Speech Recognition. *arXiv preprint*, arxiv:1610.05256, Oct. 2016.
- Z. Yuan, Z. Lyu, J. Li, X. Zhou: "An Improved Hybrid CTC-Attention Model for Speech Recognition," *arXiv preprint*, arXiv:1810.12020, Oct. 2018.

References

- S. Zhou, L. Dong, S. Xu, B. Xu: “Syllable-Based Sequence-to-Sequence Speech Recognition with the Transformer in Mandarin Chinese,” *arXiv preprint*, arxiv:1804.10752, Apr. 2018.
- A. Zeyer, P. Bahar, K. Irie, R. SchlÃ$\frac{1}{4}$ter, H. Ney: “A Comparison of Transformer and LSTM Encoder Decoder Models for ASR,” IEEE Automatic Spreech Recongnition and Understanding Workshop(ASRU), Singapore, Dec. 2019, *to appear*.
- A. Zeyer, K. Irie, R. Schlüter, H. Ney: “Improved Training of End-to-End Attention Models for Speech Recognition,” *arXiv preprint*, arXiv:1805.03294, May 2018.
- A. Zeyer, R. Schlüter, H. Ney: “Towards Online-Recognition with Deep Bidirectional LSTM Acoustic Models,” *Interspeech*, Sep. 2016.