Automatic Speech Recognition: State-of-the-Art in Transition

A Neural Paradigm Change?

Ralf Schlüter
Human Language Technology and Pattern Recognition
Lehrstuhl Informatik 6
Department of Mathematics, Computer Science and Natural Sciences
RWTH Aachen University

Preamble

- joint work with members of HLT & PR lab (Informatik 6):
  - acoustic modeling: Patrick Doetsch, Pavel Golik, Tobias Menne, Zoltan Tüske, Albert Zeyer, ...
  - language modeling: Martin Sundermeyer, Kazuki Irie, ...
  - cf. `hltpr.rwth-aachen.de/web/Publications`

- toolkits used for our own results presented here are available on our web site:
  - RASR: RWTH Automatic Speech Recognition toolkit (also handwriting)
  - RWTHLM: RWTH neural network based Language Modeling toolkit (esp. LSTM)
  - RETURNN: RWTH Extensible Training for Universal Recurrent Neural Networks (**new!**)
  - ...
  - cf. `hltpr.rwth-aachen.de/web/Software`

Outline

Outline

Human Language Technology: Overview & History

Terminology:

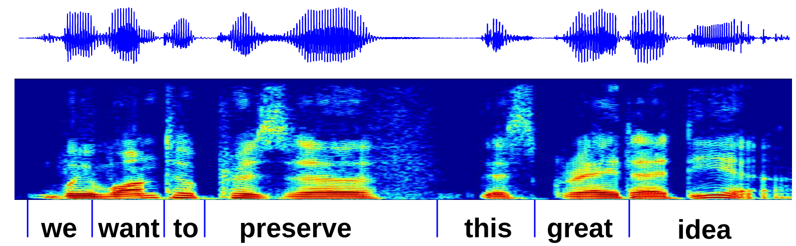
- speech: acoustic signal, spoken language
- language: text, sequence of characters, written language
- scientific disciplines:
  - NLP: natural language processing (in the strict sense): written language only
  - HLT: human language technology: spoken **and** written language

Characteristic task properties:

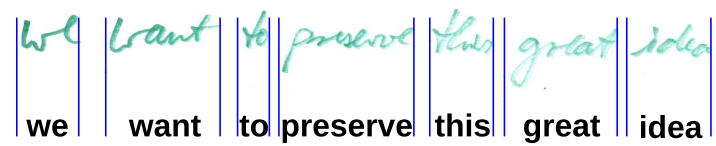
- well-defined 'classification' tasks:
  - 5000-year history of (written!) language
  - well-defined classes: letters or words of the language
- easy task for humans (at least for natives!)
- hard task for computers (as last 50 years have shown!)

Specific well-defined tasks in HLT:

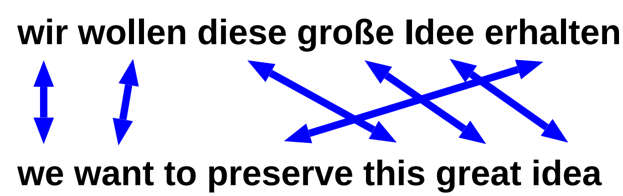
- Automatic Speech Recognition (ASR)



- Text image recognition (printed and handwritten text, offline) (HWR)



- Machine Translation (MT)



Speech and Language: Characteristic Properties

Typical situation:

input sequence $\rightarrow$ output sequence

Tasks:

- speech recognition: speech signal $\rightarrow$ words/letter sequence
- recognition of image text: (printed/written characters) text image $\rightarrow$ words/letter sequence
- machine translation: source word/letter sequence $\rightarrow$ target words/letter sequence

Common property:

output sequence = natural language word/letter sequence

Terminology:

- compound decision theory
- contextual pattern recognition
- structured output

elementary pattern classification and machine learning:

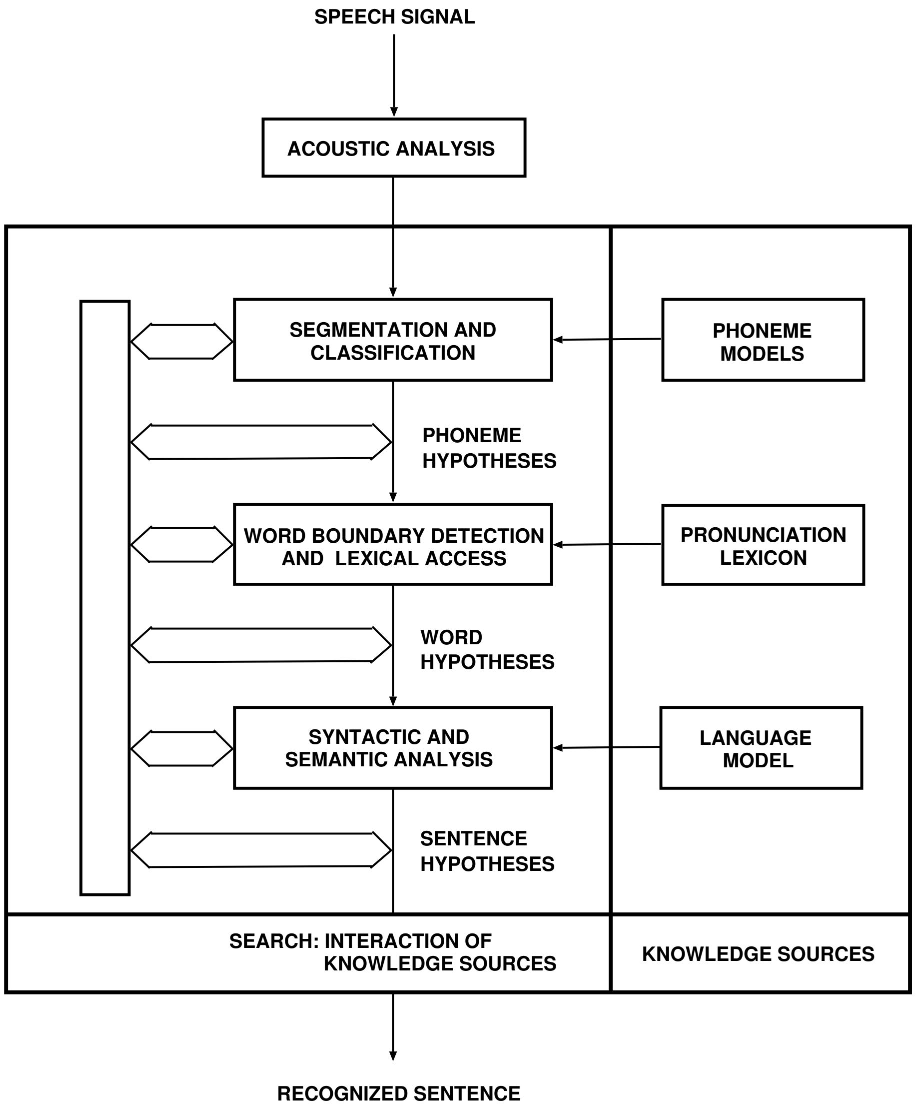
single class index without any structure

Speech recognition

What is the problem?
- ambiguities at all levels
- interdependencies of decisions

Approach [CMU and IBM 1975]:
- hypothesis scores
- probabilistic framework
- statistical decision theory

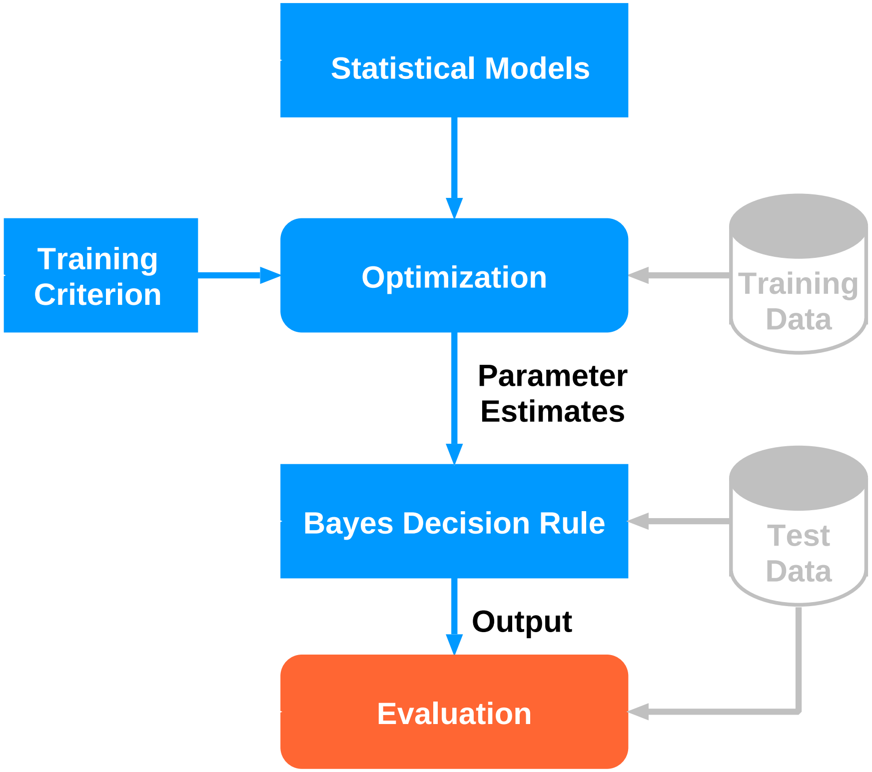
Modern terminology:
- machine learning



History Speech Recognition 1975-2015

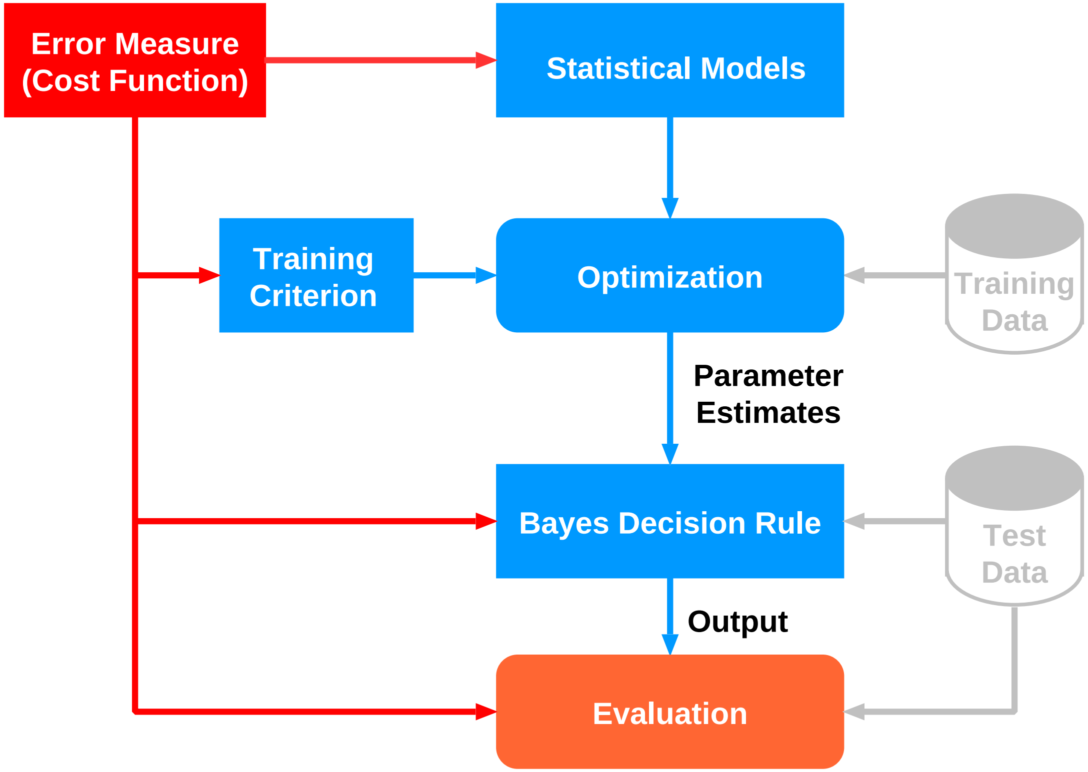
- steady increase of challenges:
  - vocabulary size: 10 digits ... 1000 ... 10.000 ... 500.000 words
  - speaking style: read speech ... colloquial/spontaneous speech
- steady improvement of statistical methods: HMM, Gaussians and mixtures, statistical trigram language model, adaptation methods, discriminative sequence training, artificial neural nets, ...
- 1985-93: criticism about statistical approach
  - too many parameters and saturation effect
  - ... 'will never work for large vocabularies' ...
- remedy(?) by rule-based approach:
  - language models (text): linguistic grammars and structures
  - phoneme models (speech): acoustic-phonetic expert systems
  - limited success for various reasons:
    huge manual effort is required!
    problem of coverage and consistency of rules
    lack of robustness
- evaluations, experimental tests:
  - the same evaluation criterion on the same test data
  - direct comparison of algorithms and systems

Bayes Architecture for Speech Recognition (and other HLT tasks)



Speech Recognition = Modeling + Statistics + Efficient Algorithms

Bayes Architecture for Speech Recognition (and other HLT tasks)



Speech Recognition = Modeling + Statistics + Efficient Algorithms
+ Performance Measure

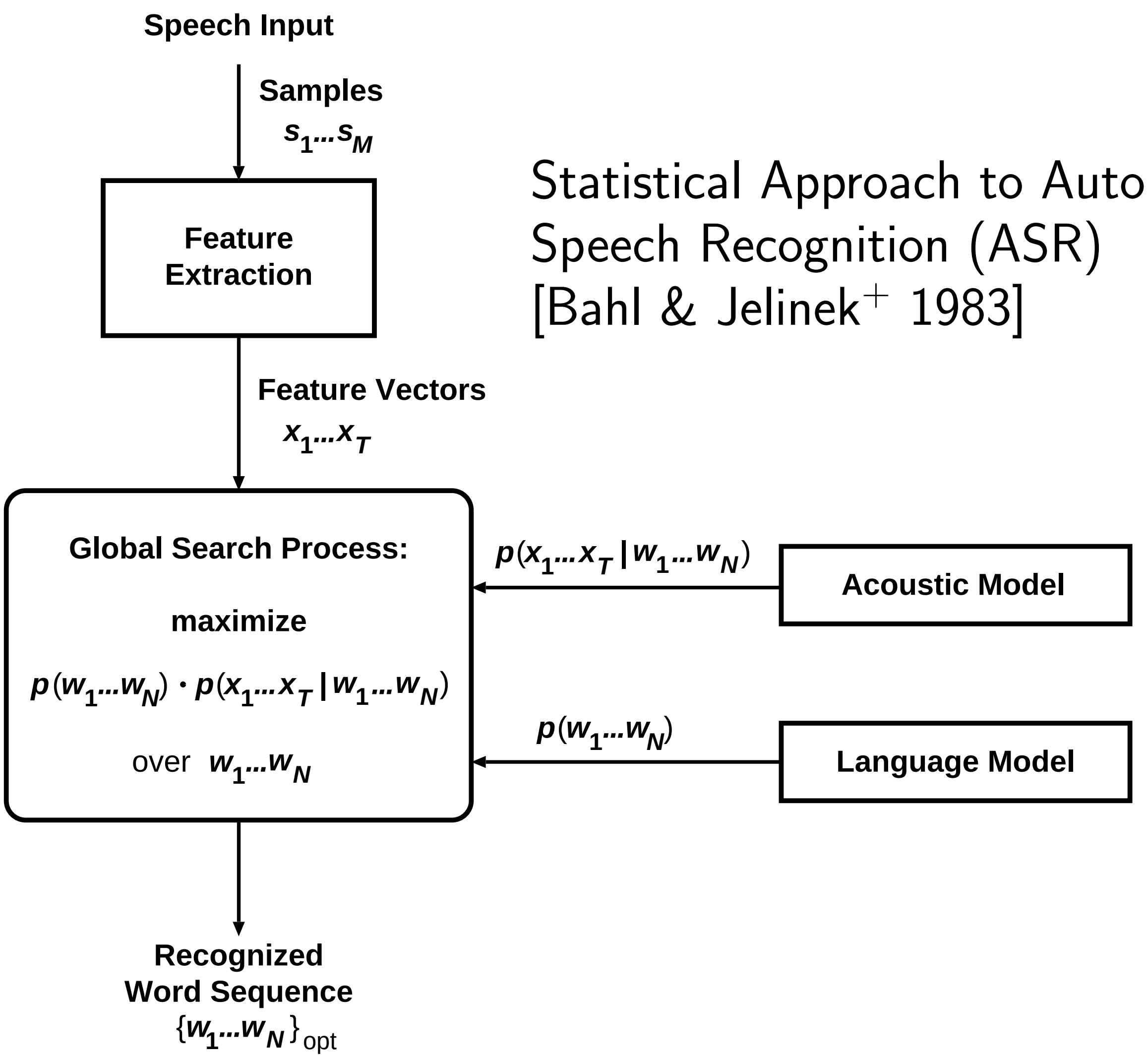
Statistical Approach

Ingredients:

- **performance measure** (often edit distance):
  to judge the quality of the system output
- **probabilistic models** (with a suitable structure):
  capture dependencies within/between input observation sequence X and output word sequence W
  - elementary observations: Gaussian mixtures, log-linear models, SVMs, NNs, ...
  - sequence context: n-gram Markov chains, HMMs, CRFs, RNNs, ...
  - effectively: discrimination function needed
- **training criterion**:
  to learn the free parameters of the models
  - ideally should be linked to performance criterion
  - might result in complex mathematical optimization (efficient algorithms!)
- **Bayes decision rule**:
  to generate the output word sequence
  - combinatorial problem (efficient algorithms)
  - should exploit structure of models

  Examples: dynamic programming and beam search, A^* and heuristic search, ...

ASR Architecture



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

Bayes Decision Rule: Sources of Errors

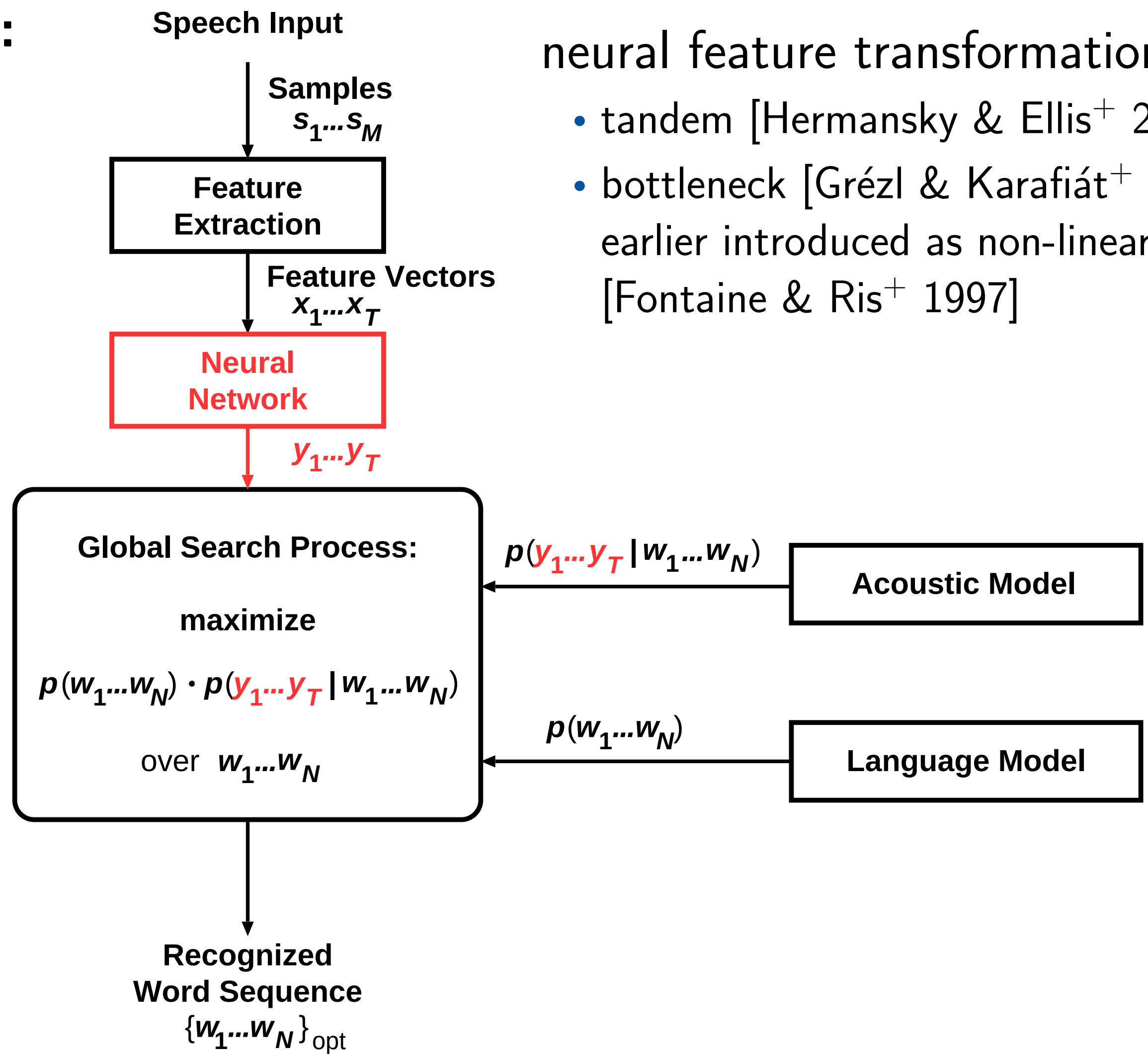
Why does a 'Bayes' decision system make errors?

To be more exact: Why errors **in addition** to so-called Bayes errors,
i.e. the minimum that can be achieved?

Reasons from the viewpoint of Bayes' decision rule:

- probability models:
  - 'incorrect' observation x: only incomplete part or poor transformation of true observations used
  - incorrect models, e.g. $p_\vartheta(c|x)$ or $p_\vartheta(c_1^N|x_1^T)$
- training conditions:
  - poor training criterion
  - not enough training data
  - mismatch conditions between training and test data
- training criterion + efficient algorithm:
  - suboptimal algorithm for training (e.g. gradient descent)
- decision rule:
  - incorrect error measure, e.g. MAP rule in ASR and MT
- decision rule + efficient algorithm:
  - suboptimal search procedure, e.g. beam search or N-best lists

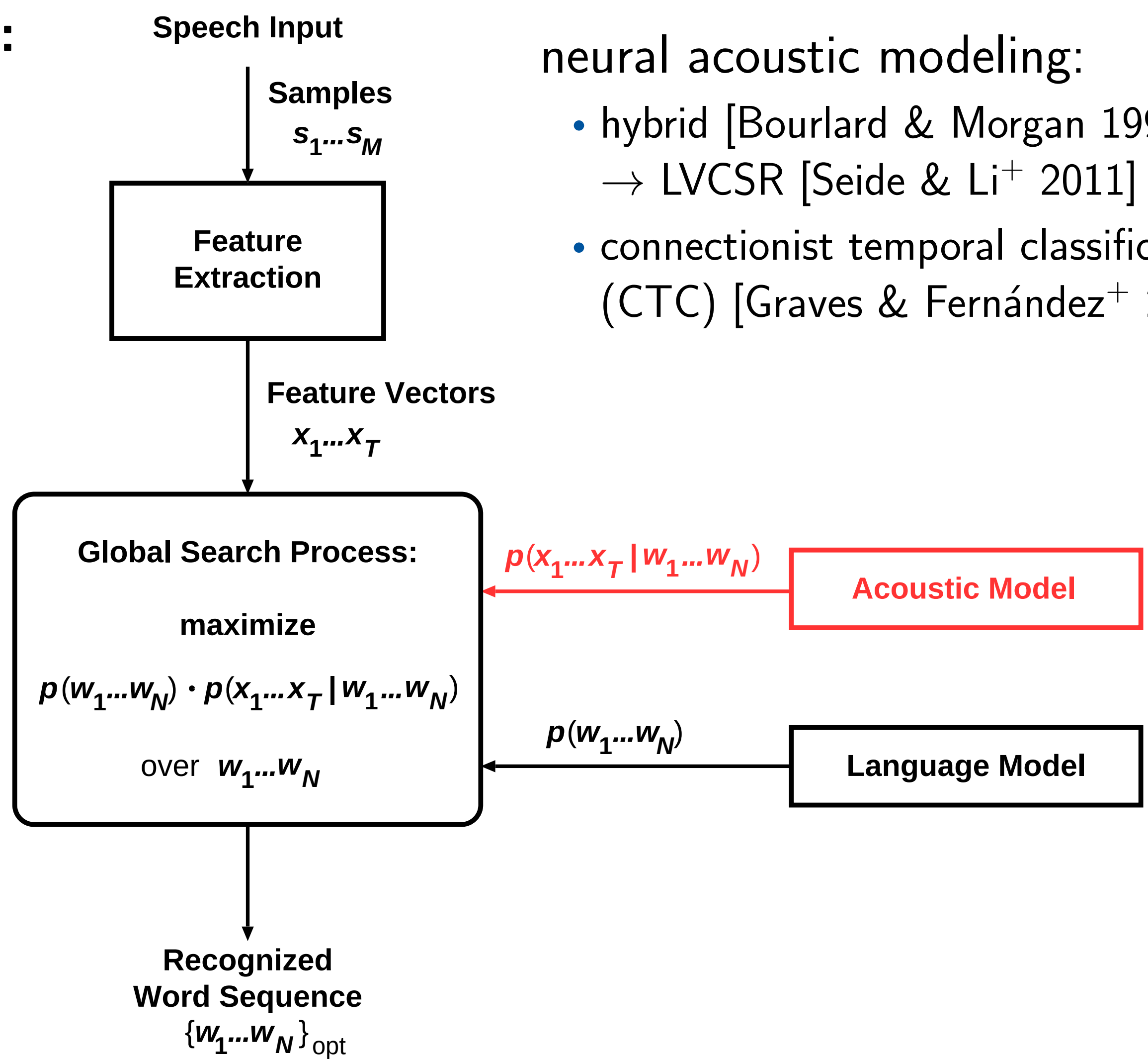
ASR Architecture: Neural Networks



neural feature transformation:

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

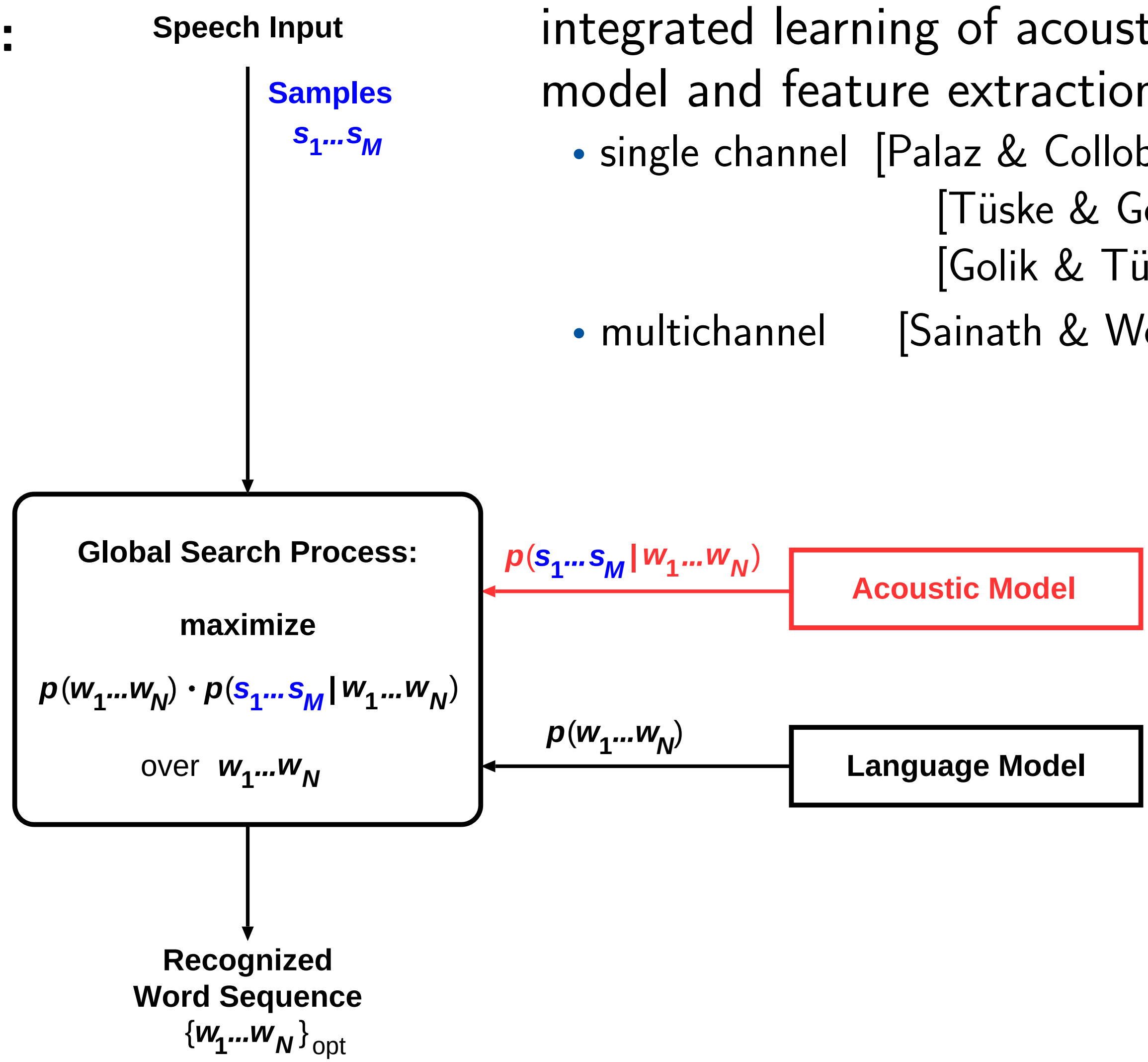
ASR Architecture: Neural Networks



neural acoustic modeling:

- hybrid [Bourlard & Morgan 1993]
  $\rightarrow$ LVCSR [Seide & Li$^+$ 2011]
- connectionist temporal classification (CTC) [Graves & Fernández$^+$ 2006]

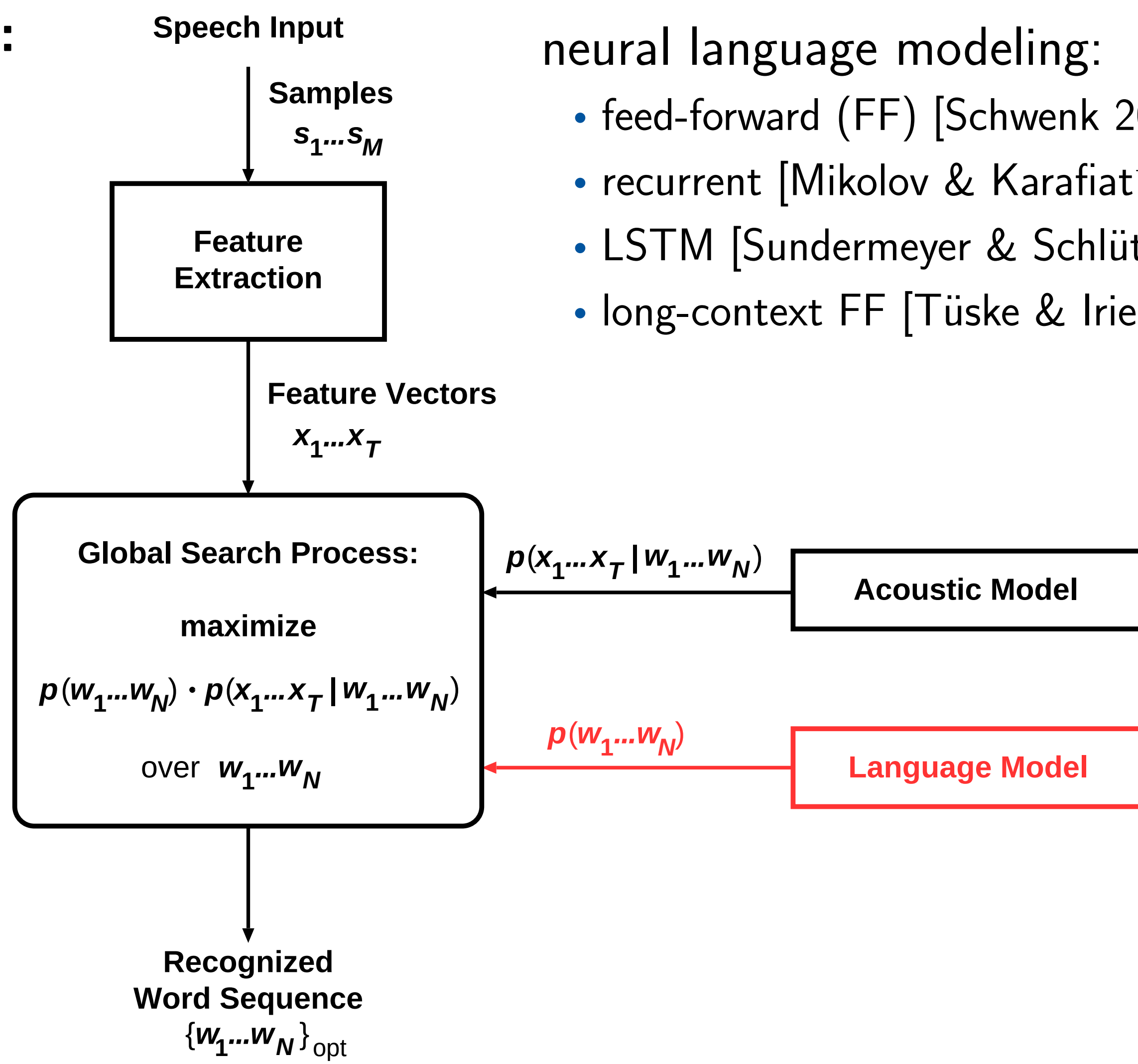
ASR Architecture: Neural Networks



integrated learning of acoustic model and feature extraction

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

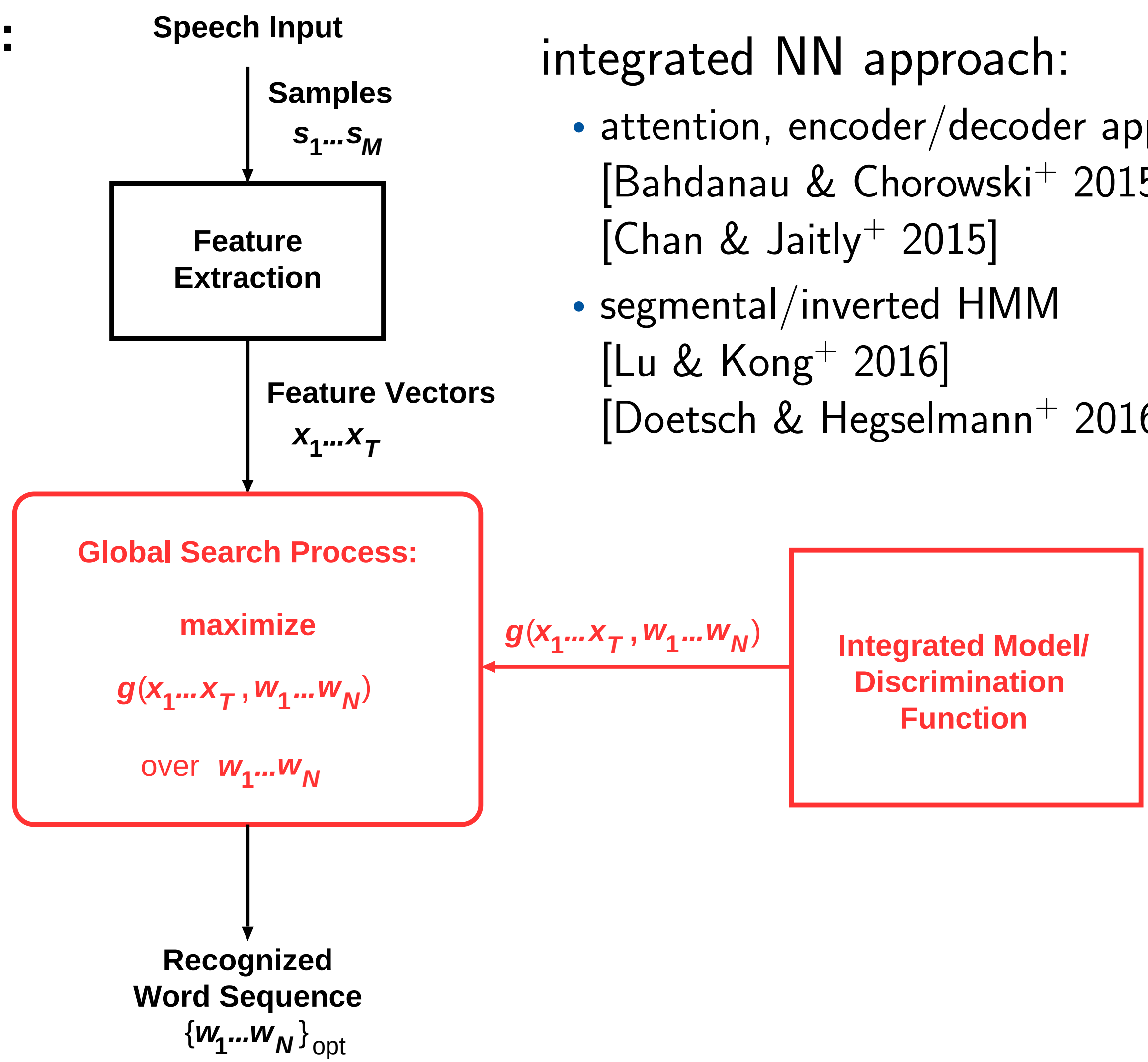
ASR Architecture: Neural Networks



neural language modeling:

- feed-forward (FF) [Schwenk 2007]
- recurrent [Mikolov & Karafiat$^+$ 2010]
- LSTM [Sundermeyer & Schlüter$^+$ 2012]
- long-context FF [Tüske & Irie$^+$ 2016]

ASR Architecture: Neural Networks



integrated NN approach:

- attention, encoder/decoder approach
  [Bahdanau & Chorowski+ 2015]
  [Chan & Jaitly+ 2015]
- segmental/inverted HMM
  [Lu & Kong+ 2016]
  [Doetsch & Hegselmann+ 2016]

Outline

Outline

Starting Points

- very complex problem: no perfect knowledge of the dependencies in speech and language:
  - different from conventional computer science
  - like a problem in natural sciences (cf. approximative modeling in physics)
- perfect solution will be difficult:
  - we accept that the system will make errors
  - but we try to find the best compromise
- fairly general view:
  - input sequence (ASR: sequence over time t: $X := x_1 \ldots x_t \ldots x_T$)
  - output sequence: $W := w_1 \ldots w_n \ldots w_N$ of unknown length N
- we need a generation mechanism:

$$X \rightarrow W = \hat{W}(X)$$

- to this purpose, we assume a
  - posterior distribution $pr(W|X)$
  - which can be extremely complex: both arguments are sequences!

Bayes Decision Rule for Sequences

- performance measure or cost function $L[\widetilde{W}, W]$ (e.g. edit distance) between true output sequence $\widetilde{W}$ and hypothesized output sequence W.
- *Bayes* decision rule minimizes expected cost:

$$X \to \overline{W}(X) := \arg\min_{W} \left\{ \sum_{\widetilde{W}} pr(\widetilde{W}|X) \cdot L[\widetilde{W}, W] \right\}$$

- standard decision rule uses sequence-level cost (MAP rule):

$$X \to \widehat{W}(X) := \arg\max_{W} \left\{ pr(W|X) \right\}$$

  since [Bahl & Jelinek$^+$ 1983], this simplified Bayes decision rule is widely used for speech recognition, handwriting recognition, machine translation, ...
  well-known inconsistency! [Jelinek 1997, pp. 4-5]
- however, standard decision rule works well, as often both decision rules agree, which can be proven under certain conditions [Schlüter & Nussbaum$^+$ 2012], e.g.:

$$L[W, \widetilde{W}] \text{ is a metric, and } \max_{W} \; pr(W|X) \geq 0.5 \quad \Rightarrow \quad \overline{W}(X) = \widehat{W}(X)$$

- approximative (second pass) sequence-level cost approaches provide good improvements [Stolcke & König$^+$ 1997, Mangu & Brill$^+$ 1999, Goel & Byrne 2000, Wessel & Schlüter$^+$ 2001]

Generative vs. Discriminative Approach

Bayes Decision Rule:

$$X \to W = \overline{W}(X) \ := \ \arg\min_{W} \Big\{ \sum_{\widetilde{W}} pr(\widetilde{W}|X) \cdot L[\widetilde{W}, W] \Big\}$$

practical considerations:

- unknown distribution $pr(W|X)$:
  remedy: replace true $pr(W|X)$ by a model $p(W|X)$
  and learn its free parameters from a HUGE set of examples
- important problem:
  - compositional modelling for $p(W|X)$ is needed since W and X are sequences
  - units smaller than the whole sequence are needed (e.g. phrases/word groups, words, letters)
- two principal approaches:
  - generative approach: $p(W, X) = p(W) \cdot p(X|W)$
    language model $p(W)$, trained on text data
    acoustic model $p(X|W)$, trained on (transcribed) audio data
  - discriminative (or direct) approach: $p(W|X) = p(W, X)/\sum_{\widetilde{W}} p(\widetilde{W}, X)$

Generative vs. Discriminative Training

Starting point:

- models $p_\theta(W)$ and $p_\theta(X|W)$ with unknown parameters θ
- training data: set of (audio, sentence) pairs (X_r, W_r), $r = 1, ..., R$

Training:

- generative model: maximum likelihood (along with EM/Viterbi algorithm):

$$F(\theta) = \sum_r \log p_\theta(W_r, X_r) = \sum_r \log p_\theta(W_r) + \sum_r \log p_\theta(X_r|W_r)$$

  nice property: decomposition into two separate problems (also: separate training data):
  - language model $p_\theta(W)$: without annotation!
  - acoustic model $p_\theta(X|W)$: with annotation!
- discriminative model: discriminative training
  - optimizes decision boundaries, e.g. maximum mutual information (MMI)
  - ideally: optim. error rate, e.g. minimum classification error (MCE), minimum phone error (MPE)
  - in practice:
    initialization by maximum likelihood
    complex optimization problem: sum over all sentences in denominator
    approximation: word lattice, many shortcuts, ...
    experiments: relative improvement by 5-10% over maximum likelihood

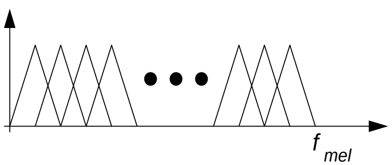
Outline

Alternative Acoustic Feature Streams

MFCC

SPEECH SIGNAL

1. PREEMPHASIS AND WINDOWING
2. MAGNITUDE SPECTRUM
3. MEL FREQUENCY WARPING: $f_{mel} = 2595\, lg(1 + \frac{f}{700\ Hz})$
4. CRITICAL BAND INTEGRATION



5. LOGARITHM
6. CEPSTRAL DECORRELATION
7. CEPSTRAL MEAN NORM.
8. ENERGY NORM.
9. SPECTRAL DYNAMIC FEATURES

ACOUSTIC VECTOR

MFCC

GT

SPEECH SIGNAL

1. PREEMPHASIS AND WINDOWING
2. GAMMATONE FILTERBANK: $h(t) = k \cdot t^{n-1} exp(-2\pi \cdot B \cdot t) \cdot cos(2\pi \cdot f_c \cdot t + \phi)$
3. RECTIFYING
4. TEMPORAL INTEGRATION
5. SPECTRAL INTEGRATION
6. 10th ROOT
7. CEPSTRAL DECORRELATION
8. CEPSTRAL MEAN NORM.
9. ENERGY NORM.
10. SPECTRAL DYNAMIC FEATURES

ACOUSTIC VECTOR

GT

PLP

SPEECH SIGNAL

1. POWER SPECTRUM
2. BARK FREQUENCY WARPING: $f_{bark} = 6\, ln(f/600 + [(f/600)^2 + 1)]^{0.5})$
3. TRAPEZOID CRITICAL BAND INTEGRATION
4. EQUAL LOUDNESS PREEMP.
5. INTENSITY-LOUDNESS POWER LAW
6. AUTOREGRESSIVE MODELING
7. LPC TO CEPSTRAL COEFF.
8. CEPSTRAL MEAN NORM.
9. ENERGY NORM.
10. SPECTRAL DYNAMIC FEATURES

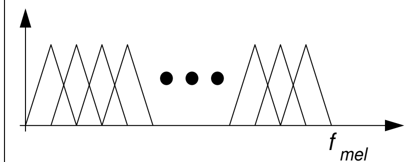
ACOUSTIC VECTOR

PLP

MF-PLP

SPEECH SIGNAL

1. POWER SPECTRUM
2. MEL FREQUENCY WARPING: $f_{mel} = 2595\, lg(1 + \frac{f}{700\ Hz})$
3. CRITICAL BAND INTEGRATION



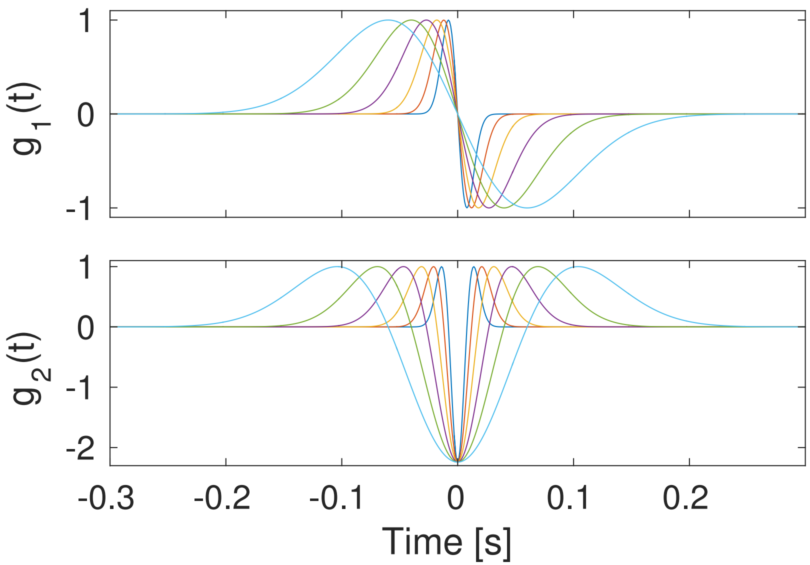
4. EQUAL LOUDNESS PREEMP.
5. INTENSITY-LOUDNESS POWER LAW
6. AUTOREGRESSIVE MODELING
7. LPC TO CEPSTRAL COEFF.
8. CEPSTRAL MEAN NORM.
9. ENERGY NORM.
10. SPECTRAL DYNAMIC FEATURES

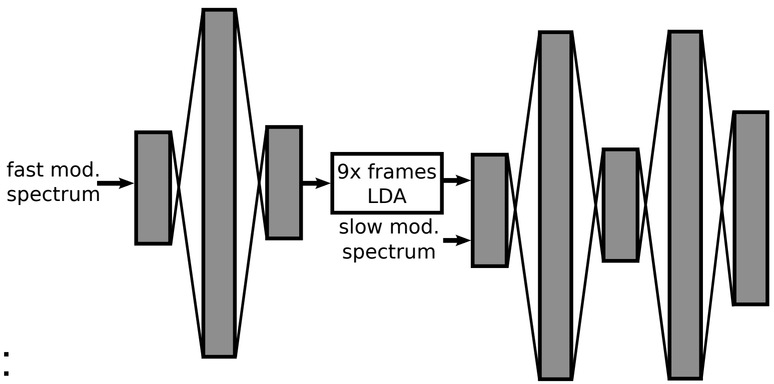
ACOUSTIC VECTOR

MF-PLP

Hierarchical MRASTA Filtering

- Long-term features:
  - Representations relAtive SpecTrA (RASTA) filtering [Hermansky & Fousek 2005].
  - Modulation frequency range (≈1-20Hz) relevant for speech perception.
- Multi-resolutional smoothing of temporal trajectories of critical band energies (CRBE)
- Filtering with first and second derivatives of Gaussians, g_1, g_2
  - σ varying in the range 8-60 ms
  - E.g. 12 temporal filters applied on 20 CRBEs + derivatives in freq.
- Processing fast and slow modulation spectrum by hierarchical MLPs





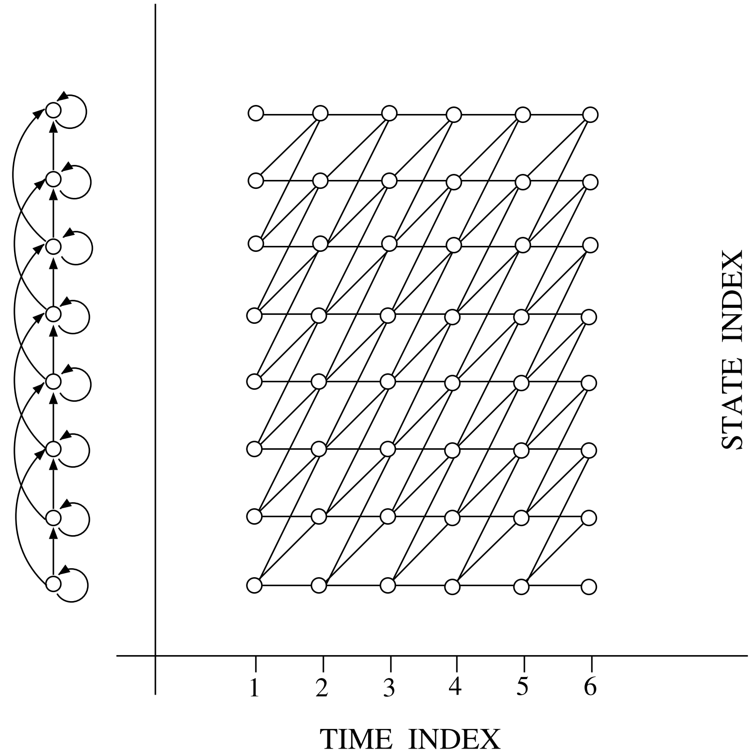
Remarks:

- FF MLPs: currently best results using MRASTA
- LSTM RNNs: filter banks sufficient, though

Outline

Speaking Rate Variation

- fundamental problem in ASR:
  variation in speaking rate,
  necessitates non-linear time alignment
- stochastic finite state machine:
  - linear chain of states $s = 1, \ldots, S$
  - transitions: forward, loop and skip
- trellis:
  - unfold over time $t = 1, \ldots, T$
  - path: state sequence $s_1^T = s_1 \ldots s_t \ldots s_T$
  - observations: $x_1^T = x_1 \ldots x_t \ldots x_T$



general view:

- two sequences without synchronization: acoustic vectors and states (with labels)
- mechanism that takes care of the synchronization (=alignment) problem

Hidden Markov Models (HMM)

The acoustic model $p(X|W)$ provides the link between sentence hypothesis W and observations sequence $X = x_1^T = x_1 ... x_t ... x_T$:

- acoustic probability $p(x_1^T|W)$ using hidden state sequences s_1^T:

$$p(x_1^T|W) = \sum_{s_1^T} p(x_1^T, s_1^T|W) = \sum_{s_1^T} \prod_t [p(s_t|s_{t-1}, W) \cdot p(x_t|s_t, W)]$$

- two types of distributions:
  - transition probability $p(s|s', W)$: not important
  - emission probability $p(x_t|s, W)$: key quantity
    realized by GMM: Gaussian mixtures models (trained by EM algorithm)
- phonetic labels (allophones, sub-phones): $(s, W) \rightarrow \alpha = \alpha_{sW}$

$$p(x_t|s, W) = p(x_t|\alpha_{sW})$$

- typical approach: models for phonemes with left and right phonetic context (triphones): decision tree (CART) clustering for finding equivalence classes
- temporal context: augment feature vector with context window around position t
- exploit first-order HMM structure for efficient search and training

Baseline HMM training:

- maximum likelihood by EM (expectation/maximization) algorithm
- looks like the ultimate and perfect solution

Positive properties:

- FULL generative model: $p_\theta(W, X) = p_\theta(W) \cdot p_\theta(X|W)$
  along with HMM for $p_\theta(X|W)$: describes the problem completely
- natural training criterion:
  - maximum likelihood, i.e. $\max_\theta \left\{ \sum_r \log\, p_\theta(W_r, X_r) \right\}$
  - virtually closed form solutions by EM algorithm
  - nice from the mathematical point of view

Negative properties:

- EM or maximum likelihood criterion
  - solves a problem that is more complex than required, i.e. $p_\theta(W, X)$ vs. $p_\theta(W|X)$
  - VERY hard from the estimation (learning) point of view
- well-known in classical pattern recognition, but ignored/overlooked in ASR:
  density estimation, i.e. learning $p_\theta(X|W)$ or $p_\theta(x_t|\alpha)$, is much harder than
  classification, i.e. learning $p_\theta(W|X)$ or $p_\theta(\alpha|x_t)$

Outline

Statistical Modeling of Syntax and Semantics

Definition of a language model (LM):

- $p(w_1^N)$: (prior) probability of the word sequence $w_1^N := w_1 ... w_n ... w_N$

Need for language model in Bayes decision rule in ASR (also SMT!):

$$x_1^T \rightarrow \hat{w}_1^{\hat{N}}(x_1^T) \;=\; \underset{N, w_1^N}{\operatorname{argmax}} \left\{ p(w_1^N) \cdot p(x_1^T | w_1^N) \right\}$$

Observations about the language model $p(w_1^N)$:

- it can be learned from text only (unlabeled data!)
- it can improve performance dramatically

Perplexity:

- quality measure for LM (based on text data, i.e. w/o a recognition experiment)
- geometric average of probability per word by computing N-th root:

$$PP \;:=\; \Big(p(w_1^N)\Big)^{-1/N} \;=\; \Big(\prod_{n=1}^{N} p(w_n | w_1^{n-1})\Big)^{-1/N}$$

define w_1^0 as empty sequence

- geometric average of inverse probability $\rightarrow$ interpretation: average effective vocabulary size

Markov Chain, Count Models

Conventional approach:

- assume Markov chain of order k:
  limit the dependence on the full history w_0^{n-1} to the immediate k predecessor words:

$$p(w_n|w_0^{n-1}) \;:=\; p_\vartheta(w_n|w_{n-k}^{n-1})$$

  terminology: $(k+1)$-gram, e.g. four-, tri-, bi-, unigram (w_n^{n-1} defines empty context for unigram)
- free parameters ϑ to be learned from training data:
  conditional probabilities $p_\vartheta(w_n|w_{n-k}^{n-1})$ for the $(k+1)$-gram events
- natural training criterion for a corpus w_1^N: minimum perplexity

$$\max_\vartheta \left\{ \frac{1}{N} \sum_{n=1}^{N} \log p_\vartheta(w_n|w_{n-k}^{n-1}) \right\} \xrightarrow{N\to\infty} \max_\vartheta \left\{ \sum_{w,h_1^k} pr(w|h_1^k) \cdot \log p_\vartheta(w|h_1^k) \right\}$$

  - equivalent to cross-entropy training (or maximum likelihood)
  - resulting estimates: relative frequencies based on event counts

Unseen Events, Smoothing

Problem:

- most of the events are never seen in training data
- example: vocabulary of $100k = 10^5$ words results in 10^{15} possible trigrams
- result: virtually all event counts are zero

Remedy:

- interpolation/combination of LMs of various orders k,
  e.g. fivegrams, fourgram, trigram, bigram and unigram events
- various strategies:
  - models: interpolation or back-off
  - estimation: cross-validation or leave-one-out
  - concept of generalized marginal distributions, e.g. going from trigrams to bigrams
- most strategies implemented in LM toolkit by SRI

Outline

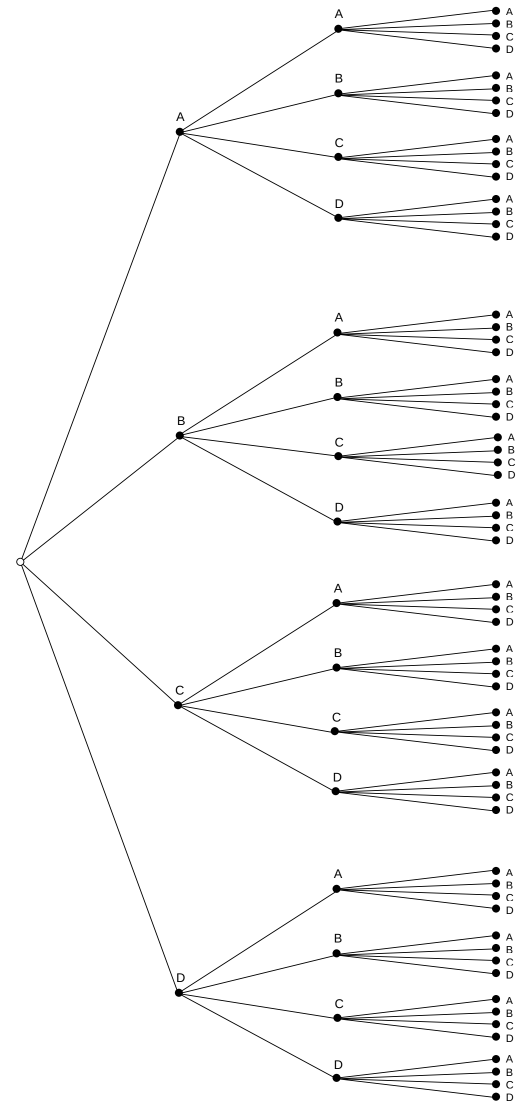
Search Space

Combinatorial complexity

- *Bayes* decision rule involves optimization over all possible word sequences and alignments
- Number of word sequences and number of alignment paths rise exponential with length

Dynamic programming

- Markov assumptions in HMM and LM can be exploited for efficient search
- Recursion equations reduce complexity to being linear in input length and polynomial in vocabulary size
- For limited vocabularies and LM context **exact** solution of optimization problem possible.



Beam Search

Large vocabulary

- even for moderate LM context, for large vocabularies ($\gtrsim$ 10k), exhaustive search becomes prohibitive
- **approximations** are needed for efficient search
- utilize probabilistic scoring for hypothesis pruning

Dynamic programming hypothesis pruning

- time-synchronous propagation of partial dynamic programming hypotheses
- discard hypotheses relative to current best hypotheses
- goal: complexity overall linear in input

Interrelation with Modeling

- more sophisticated models usually introduce higher complexity into system
- **however**: scores become more pronounced
- allows for tighter pruning, compensates increase in complexity

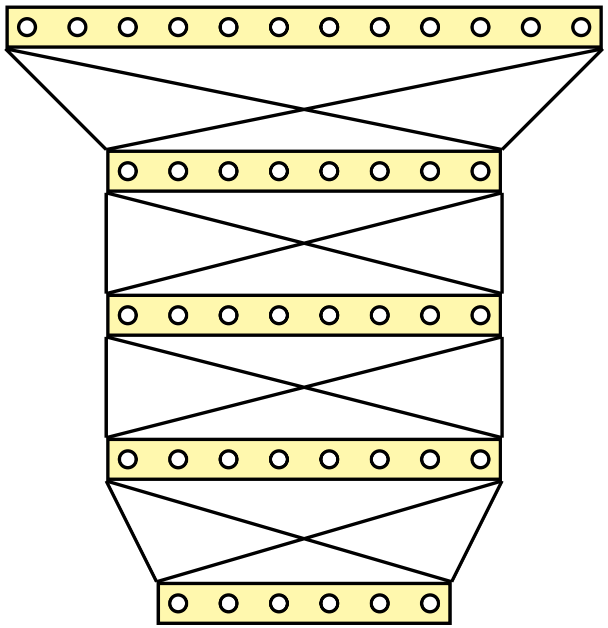
Outline

Outline

(First) NN Renaissance around 1986

Various interpretations/justifications:

- human/biological brain
- massive parallelism
- mathematical viewpoint:
  modelling ANY input-output relation



Typical ANN structure:

- MLP: feedforward multi-layer perceptron
- with input, hidden and output layers

Theoretical results:

- one hidden layer should be sufficient (!?)
  [Cybenko 1989, Hornik & Stinchcombe^{+} 1989]

Training:

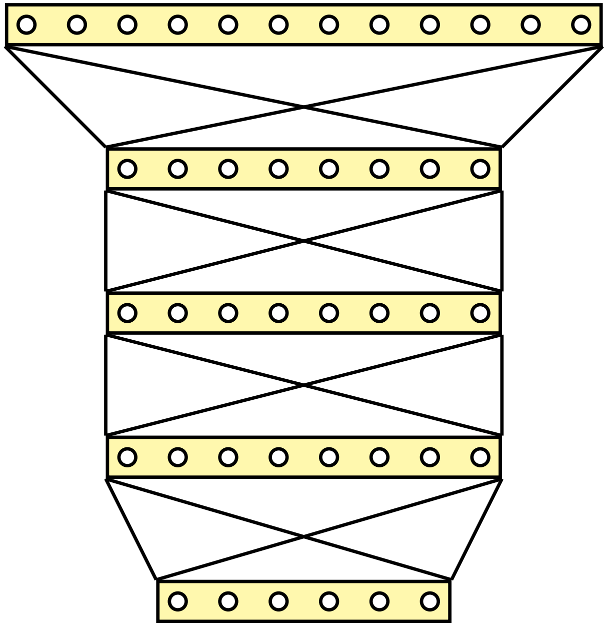
- (hard) optimization problem with millions of free parameters (= weights)

Classical Architecture:

Feedforward Multi-Layer Percerptron (FF-MLP)

- task: classification with observation vector $x \in \mathbb{R}^D$ and associated class c

Architecture:



- several layers (feedforward links only, no recurrence)
- input layer = observation vector x:
  each node represents a vector component
- between layers:
  - matrix-vector product for layer pair
  - nonlinear activation function
- output layer:
  - softmax normalization
  - each output node represents a class c and its associated score $p_\vartheta(c, x)$
- set ϑ of all weights (parameters) of the FF-MLP

ANN Activation Functions

Examples of activation functions:

- sigmoid function (also called logistic function):

$$u \to \sigma(u) = \frac{1}{1 + \exp(-u)} \qquad \in [0, 1]$$

- hyperbolic tangent:

$$u \to \tanh(u) \;=\; 2\,\sigma(2u) - 1 \qquad \in [-1, 1]$$

  - in principle: no difference to sigmoid $\sigma(\cdot)$
  - in practice: difference due to side effects

- rectifying linear unit: $u \to r(u) = \max\{0, u\}$
  - so far: not useful in symbolic processing (?)
- softmax function:

$$u_c \to S(u_c) = \frac{\exp(u_c)}{\sum_{\tilde{c}} \exp(u_{\tilde{c}})} \qquad \text{with} \quad \sum_c S(u_c) = 1.0$$

  - generates normalized output for (probability distribution over) each node c of the layer under consideration (typically: output layer)

Outline

Classification with Artificial Neural Networks

Decision rule for observation (vector) x:

$$x \rightarrow \hat{c}_x := \operatorname{argmax}_c \left\{ p_\vartheta(c, x) \right\}$$

Ideal values at output nodes:

- correct class: 1
- wrong class: 0

Distinguish varying conditions for decision rule:

- no context, in isolation (here)
- context of a sequence (see later)

Training criteria:

- squared error: unconstrained output: $p_\vartheta(c, x) \in \mathbb{R}$

$$F_{\text{SE}}(\vartheta) := \frac{1}{N} \sum_{n=1}^{N} \sum_{c} [p_\vartheta(c, x_n) - \delta(c, c_n)]^2$$

- cross-entropy: normalized output: $p_\vartheta(c, x) \in [0, 1]$: $\quad \sum_c p_\vartheta(c, x) = 1$

$$F_{\text{CE}}(\vartheta) := \frac{1}{N} \sum_{n=1}^{N} \log \; p_\vartheta(c_n | x_n)$$

Training Criteria: Interpretation & Relation to Error Rate

Straightforward analysis shows important result for both training criteria:

- ANN outputs are (estimates of) true **class posterior probabilities**!
- result independent of any training strategy (e.g. type of backpropagation)
- assumes sufficient flexibility and parameters in ANN
- generalization capability from training to test set: not addressed

Gradient search (backpropagation):

- we can only find a **local** optimum
- there may be a huge number of local optima; but most of them seem to be equivalent
- experimental evidence: backpropagation able to find local optimum that's typically 'good enough'
- generalization capability: implicitly taken into account by cross-validation (early stopping) ?

Relation between error rate and training criteria?

- we need a strict distinction:
  - error rate for the true distribution: Bayes classification error
  - error rate for the learned distribution: model classification error
- training criteria: tight upper bound for squared difference between these two error rates [Ney 2003]
- **remark**: this result does *not* address the generalization problem

Outline

Conventional view: consider MLP with softmax output

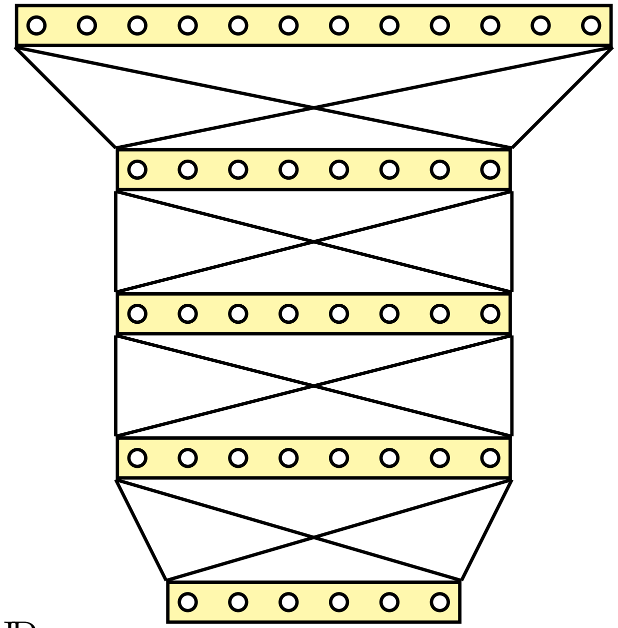
- input layer: raw input vector z
- hidden layers perform feature extraction:

$$x = f(z)$$

  with feature vector $x \in \mathbb{R}^D$ before output layer
  note: no dependence on class labels $c = 1, ..., C$
- output layer: probability distribution over classes c

$$p(c|x) = \frac{\exp(\lambda_c^T \cdot x + \gamma_c)}{\sum_{c'} \exp(\lambda_{c'}^T \cdot x + \gamma_{c'})}$$

  with output layer weights $\lambda_c \in \mathbb{R}^D$ and offsets (biases) $\gamma_c \in \mathbb{R}$



Interpretation of MLP with softmax output:

- **feature extraction** followed by a **log-linear classifier**

Relation to generative modeling [Heigold & Schlüter$^+$ 2012]:

- softmax operation results from using class posterior distribution of a Gaussian model

Outline

Sequence Processing

So far:

- handling of (input, output) pairs (c, x) in isolation
- no internal structure in c or x (unlike sequences)

From single events to sequences:

- consider a pair of synchronized input and output sequence over time t:

$$(c_t, x_t), \ t = 1, ..., T$$

  with input vectors x_t and class labels c_t
- goal: model the conditional probability $p(c_1^T | x_1^T)$ of the sequence c_1^T (assuming causality and a special start symbol c_0):

$$p(c_1^T | x_1^T) = \prod_t p(c_t | ...)$$

  with ANN output vector $y_t = p(c_t | ...)$ at each time t

Sequences with Synchronisation

Illustration:

- model with 1:1 correspondence between class labels c_1^T and observations x_1^T
- sequence length T is known

observations x_1^T:	x_1	x_2	...	x_{t-1}	x_t	x_{t+1}	...	x_{T-1}	x_T
	|	|	|	|	|	|	|	|	|
class labels x_1^T:	c_1	c_2	...	c_{t-1}	c_t	c_{t+1}	...	c_{T-1}	c_T

typical problems:

- spelling correction (character level)
- POS tagging (POS: parts of speech)
- frame labelling in ASR (incl. pronunciation and language models!) and acoustic scores in hybrid HMMs
- recognition problems with no problems of boundary detection: isolated words, printed character recognition, ...

Factorization of Conditional Probability $p(c_1^T|x_1^T)$

- conditional independence in c_1^T with look-ahead for x_1^T: $p(c_1^T|x_1^T) = \prod_{t=1}^T p_t\left(c_t|x_1^T\right)$

observations x_1^T:	x_1	x_2	...	x_{t-1}	x_t	x_{t+1}	...	x_{T-1}	x_T
	|	|	|	|	|	|	|	|	|
class labels c_1^T:	—	—	...	—	$\boxed{c_t}$	—	...	—	—

- conditional dependence in c_1^T without look-ahead in x_1^T: $p(c_1^T|x_1^T) = \prod_{t=1}^T p\left(c_t|c_0^{t-1}, x_1^t\right)$

observations x_1^T:	x_1	x_2	...	x_{t-1}	x_t	—	...	—	—
	|	|	|	|	|	|	|	|	|
class labels c_1^T:	c_1	c_2	...	c_{t-1}	$\boxed{c_t}$	—	...	—	—

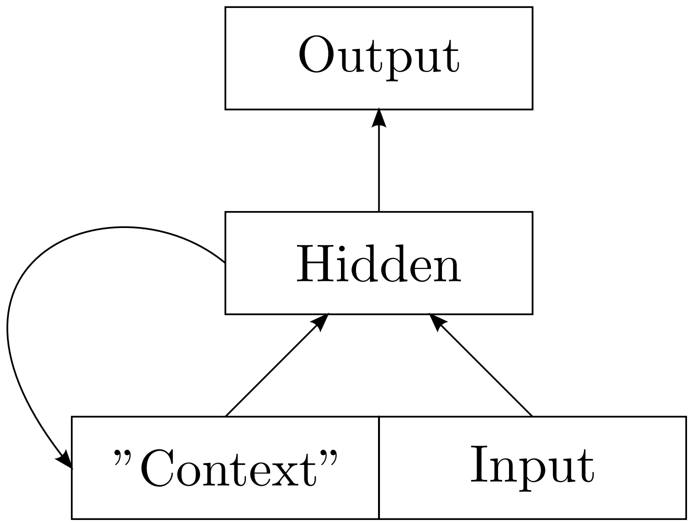
- conditional dependence in c_1^T with look-ahead in x_1^T: $p(c_1^T|x_1^T) = \prod_{t=1}^T p\left(c_t|c_0^{t-1}, x_1^T\right)$

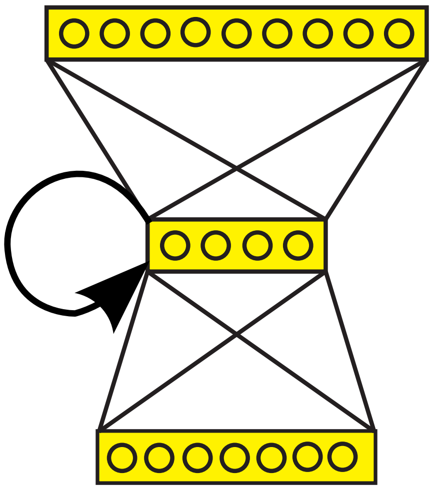
observations x_1^T:	x_1	x_2	...	x_{t-1}	x_t	x_{t+1}	...	x_{T-1}	x_T
	|	|	|	|	|	|	|	|	|
class labels c_1^T:	c_1	c_2	...	c_{t-1}	$\boxed{c_t}$	—	...	—	—

Recurrent Neural Network (RNN): Principle

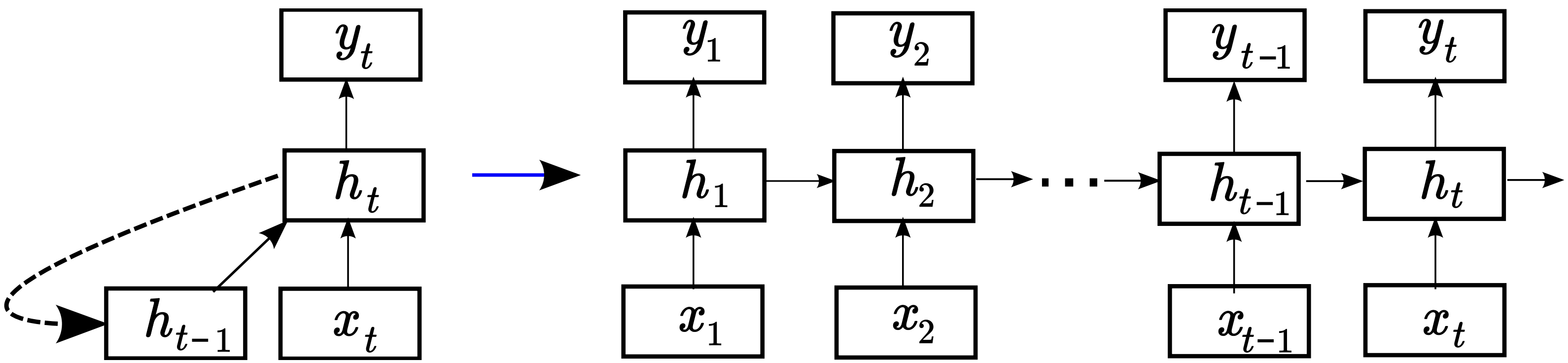
principle:

- introduce a **memory** (or context) component to keep track of history
- result: there are two types of input: memory h_{t-1} and observation x_t





Unfolding RNN over Time



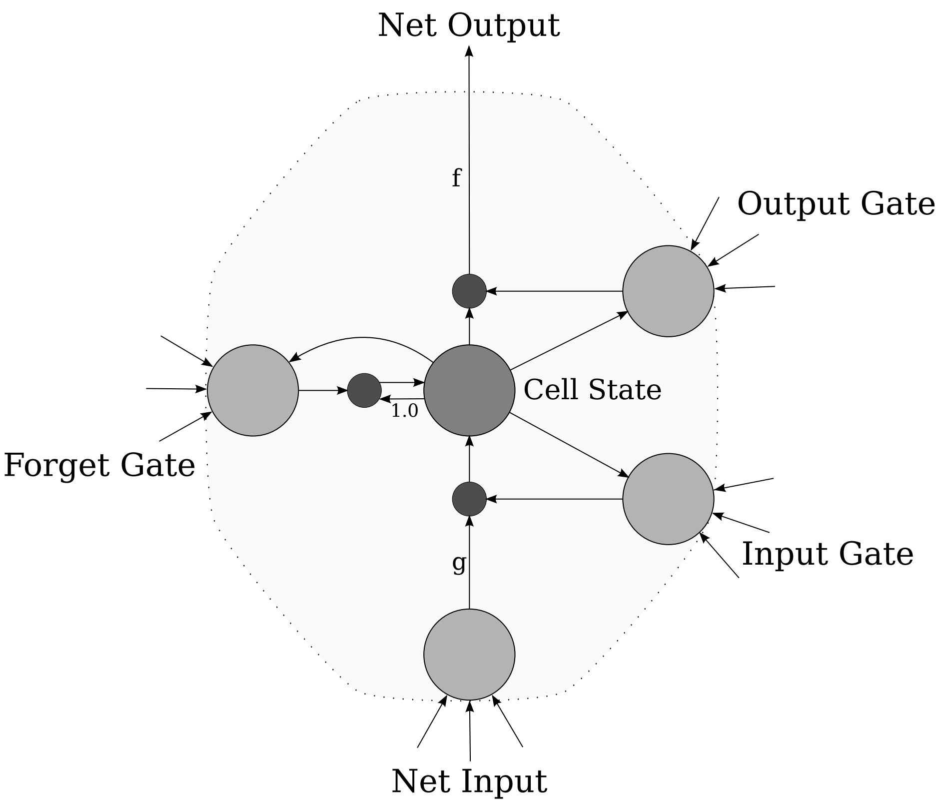
The architecture of RNN can be unfolded over time:

- We get a feedforward network with a special **deep** architecture.
- The application of the backpropagation algorithm to this unfolded network is called **backpropagation through time**.

LSTM RNN [Hochreiter & Schmidhuber 1997, Gers & Schraudolph^{+} 2002]

extension of (simple) RNN by
LSTM: long short-term memory

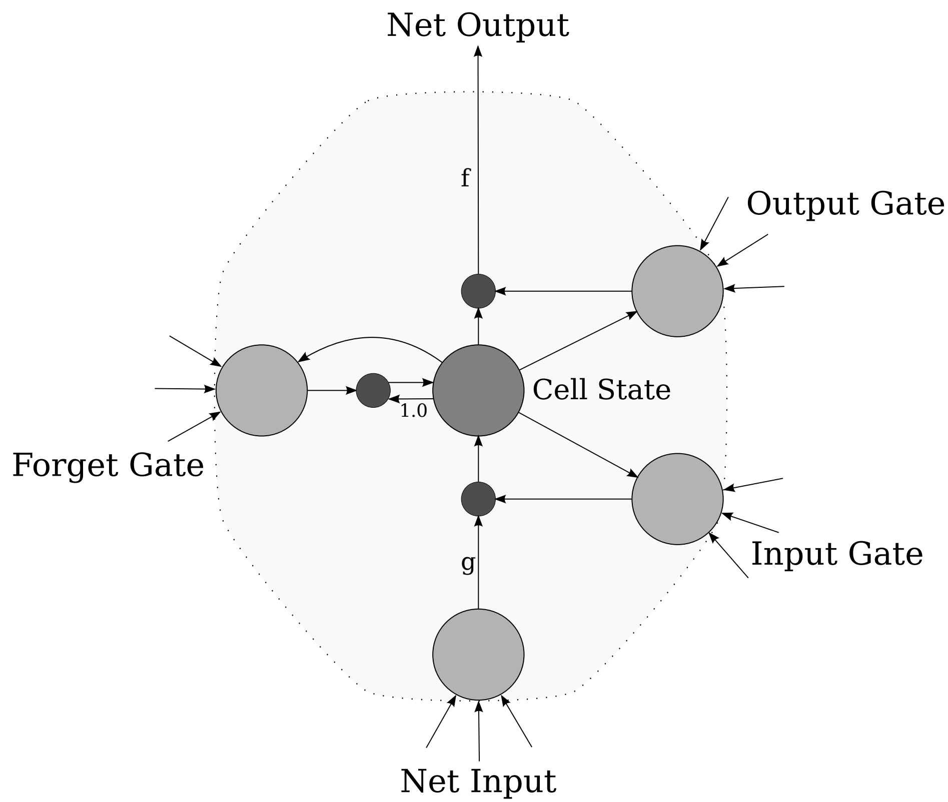
- problems of simple RNN:
  - vanishing/exploding gradients
  - no protection of memory h_t
- remedy by LSTM architecture:
  control the access to its internal memory
  by introducing gates/switches
- refinements:
  - bidirectional structure
  - several hidden layers



LSTM RNN [Hochreiter & Schmidhuber 1997, Gers & Schraudolph+ 2002]

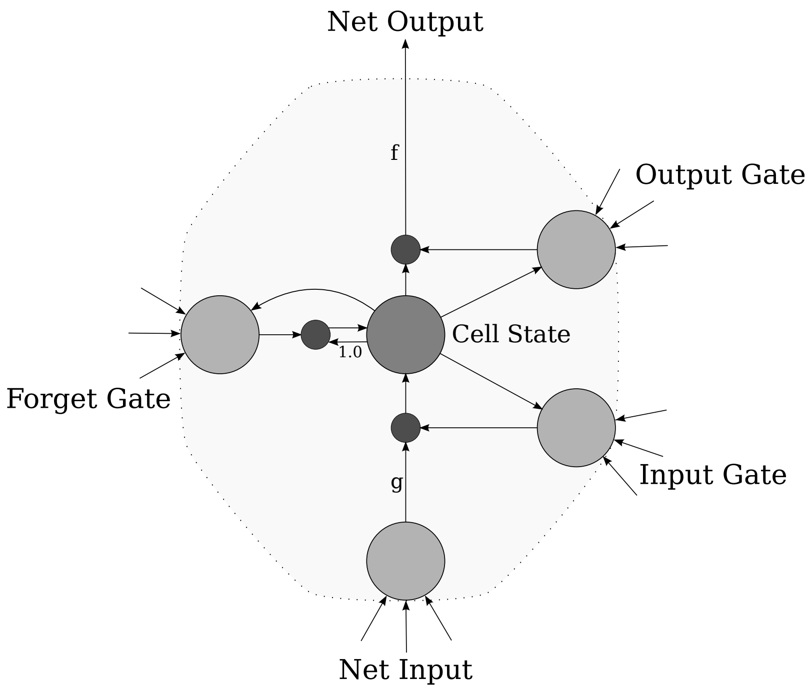
LSTM approach:

- split RNN hidden vector h_t into (memory) cell state c_t and net output s_t
- overall LSTM operations involve three 'input' vectors at time t: s_{t-1}, c_{t-1}, x_t
- update operations at time t:
  cell state: $c_t = c_t(s_{t-1}, c_{t-1}, x_t)$
  net output: $s_t = s_t(s_{t-1}, c_{t-1}, x_t)$
  output layer: $y_t = y_t(s_t)$ with softmax
- introduce three gates (input, output, forget) to control the information flow



LSTM Architecture

- three vectors (over time t): c_t, s_t, x_t
- gates (or switches): use sigmoid function $\sigma(\cdot)$
- full matrices $(A_2, R; A_i, R_i, A_f, R_f, A_o, R_o)$ and diagonal matrices (W_i, W_f, W_o)
- usual matrix and vector operations and element-wise multiplication $\odot$
- Net Input (like update formula of simple RNN):
$$z_t = \tanh(A_2 x_t + R s_{t-1})$$
- Should this Net Input z_t access the Cell State c_t?
Input Gate: $i_t = \sigma(A_i x_t + R_i s_{t-1} + W_i c_{t-1})$
- Should the Cell State c_{t-1} be forgotten?
Forget Gate: $f_t = \sigma(A_f x_t + R_f s_{t-1} + W_f c_{t-1})$
- Based on i_t and f_t, update the Cell State c_t:
$$c_t = f_t \odot c_{t-1} + i_t \odot z_t$$
- Should this update c_t be output?
Output Gate: $o_t = \sigma(A_o x_t + R_o s_{t-1} + W_o c_t)$
- Based on o_t, compute the Net Output:
$$s_t = o_t \odot c_t$$



RNN and probabilities: What does a general RNN compute?

note: general RNN includes LSTM as a special case
two sequences over time $t = 1, ..., T$:

$$\begin{aligned} \text{input: sequence of observations:} \quad x_1^T &= x_1 ... x_t ... x_T \\ \text{output: sequence of class labels:} \quad c_1^T &= c_1 ... c_t ... c_T \end{aligned}$$

consider the posterior probabilty of the output sequence:

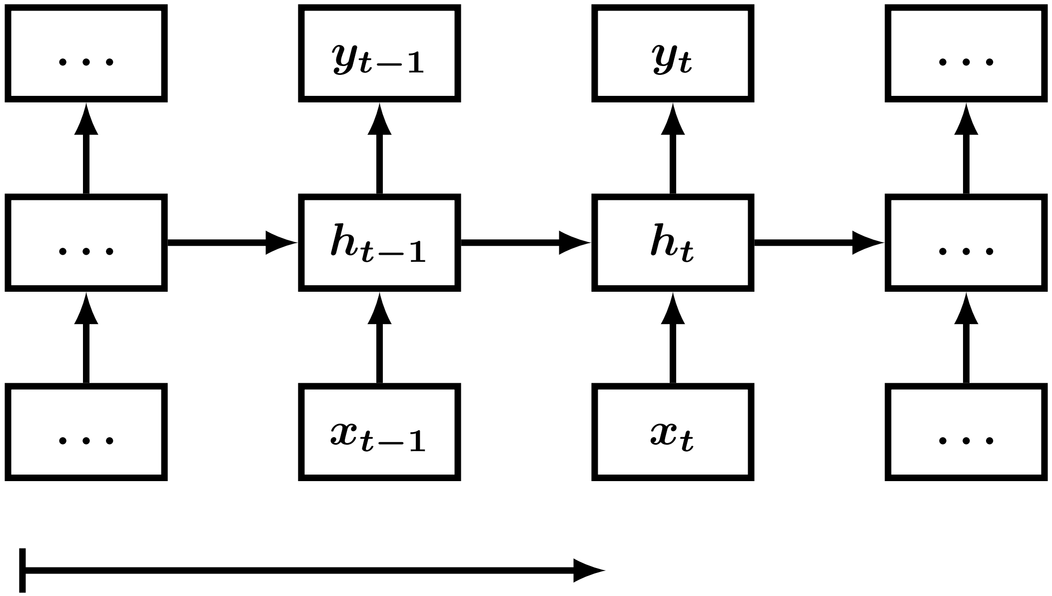
$$\begin{aligned} \text{factorization over time } t: \qquad p(c_1^T | x_1^T) &= \prod_{t=1}^{T} p(c_t | c_0^{t-1}, x_1^T) \\ \text{marginalization for time } t: \qquad \sum_{c_1^T : c_t = c} p(c_1^T | x_1^T) &= p_t(c | x_1^T) \\ \text{more ...} \qquad & \end{aligned}$$

notation for RNN output vector with nodes = classes $c = 1, ..., C$:

$$y_t = [y_t(c)] = [p_t(c | ...)]$$

RNN: Variant 1

uni-directional, no feedback of output labels

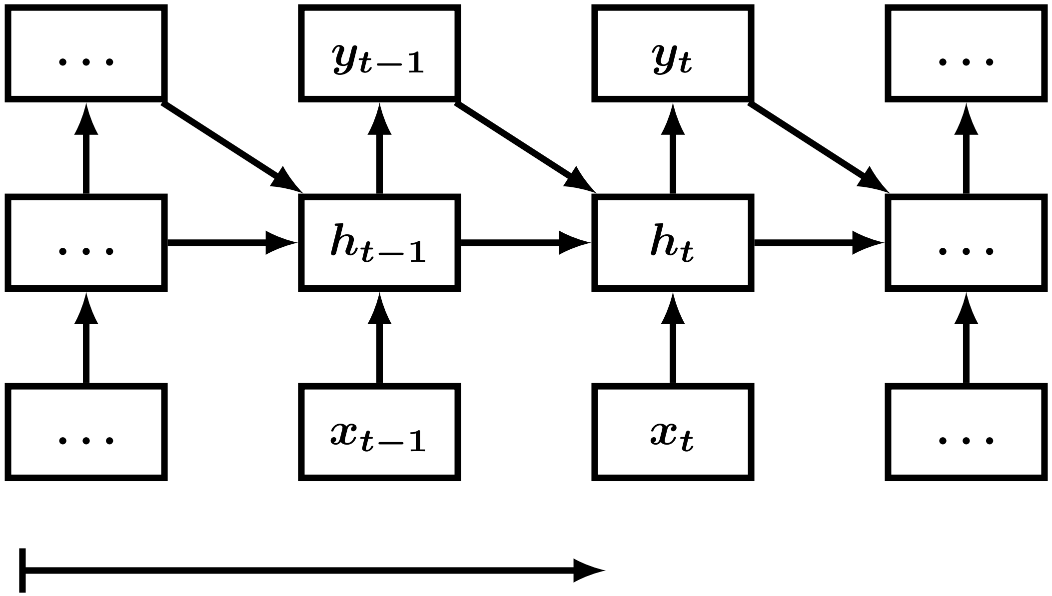


RNN output vector:

$$y_t(c) = p_t(c|x_1^t)$$

RNN: Variant 2

uni-directional, with feedback of output labels

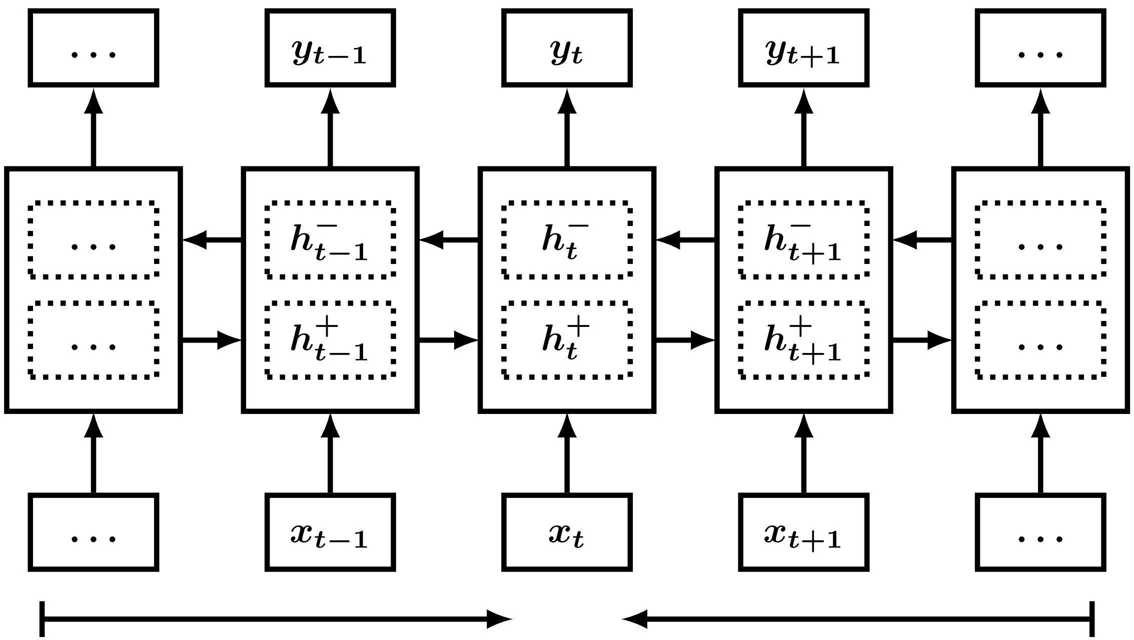


RNN output vector:

$$y_t(c) = p_t(c|c_0^{t-1}, x_1^t)$$

RNN: Variant 3

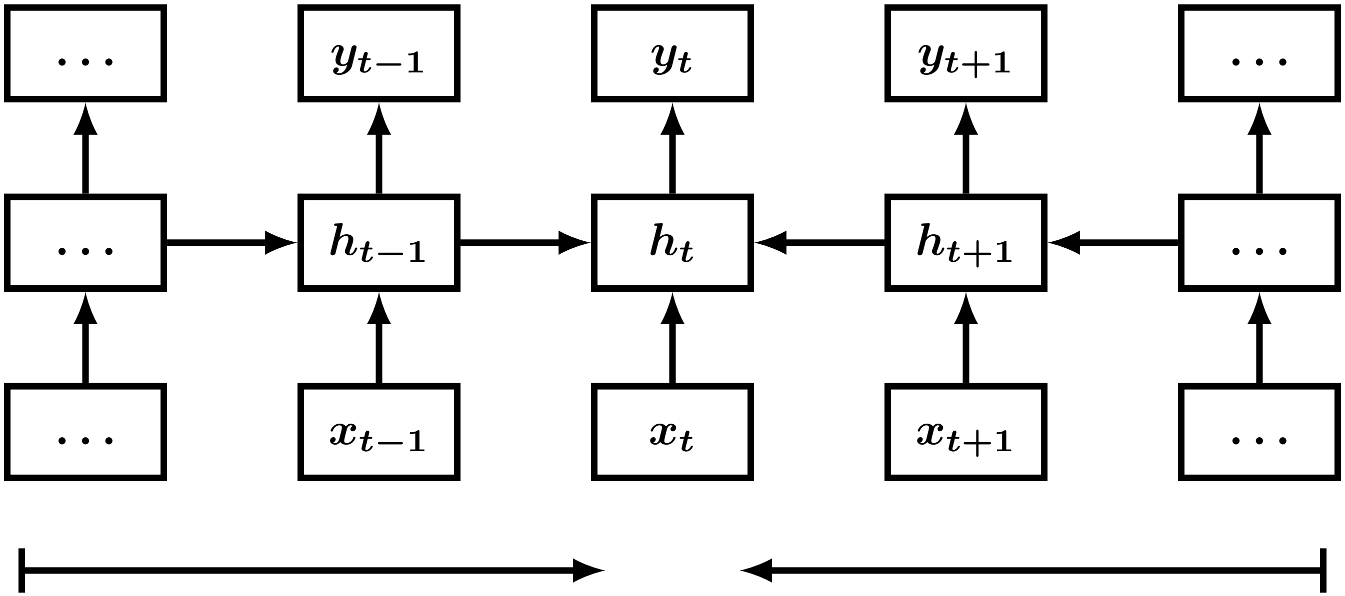
bi-directional, no feedback of output label



Internal Structure: Separate Forward and Backward Hidden Layers

RNN: Variant 3

bi-directional, no feedback of output label

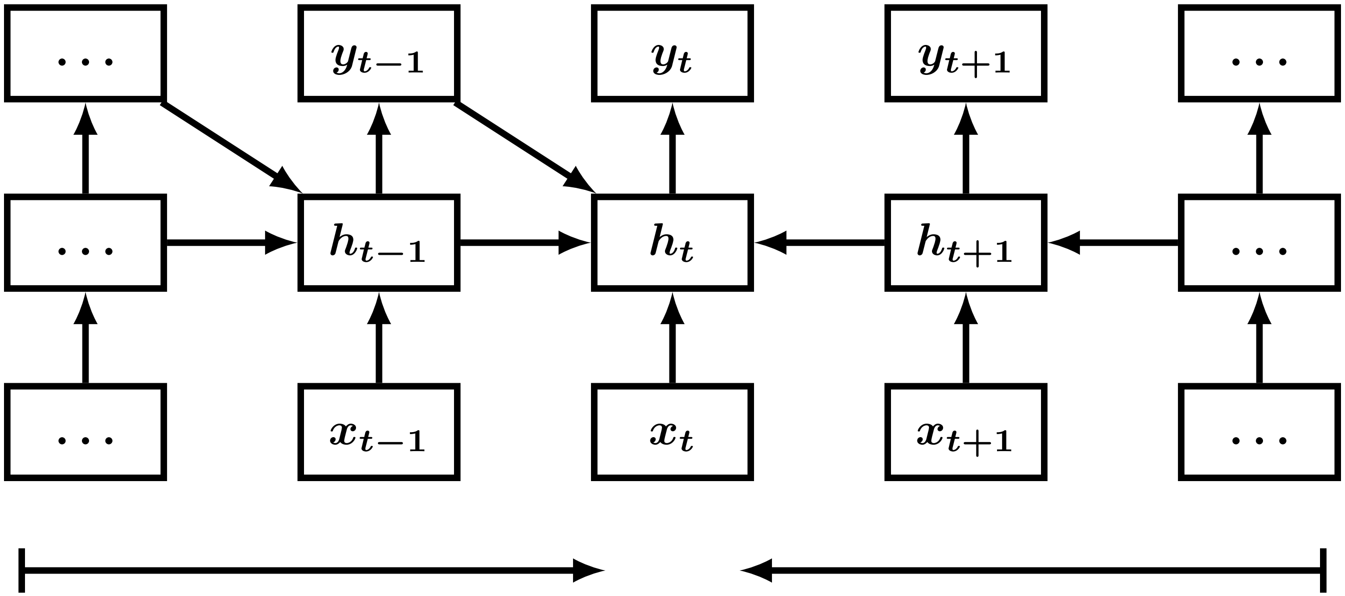


RNN output vector:

$$y_t(c) = p_t(c|x_1^T)$$

RNN: Variant 4

bi-directional, with uni-directional feedback of output label

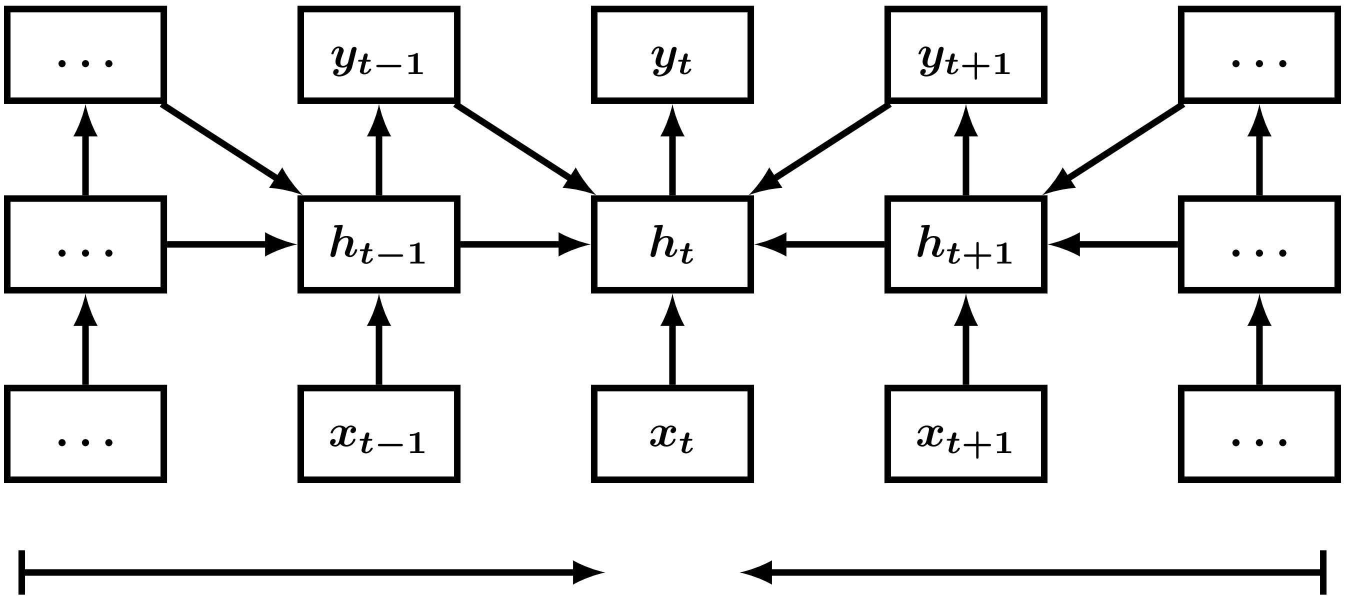


RNN output vector:

$$y_t(c) = p_t(c|c_0^{t-1}, x_1^T)$$

RNN: Variant 5

bi-directional, with bi-directional feedback of output label



RNN output vector:

$$y_t(c) = p_t(c|c_0^{t-1}, c_{t+1}^T, x_1^T)$$

Overview of RNN Outputs

label feedback	no	uni-direct.	bi-direct.
uni-dir. RNN	$p_t(c\|x_1^t)$	$p_t(c\|c_0^{t-1}, x_1^t)$	——
bi-dir. RNN	$p_t(c\|x_1^T)$	$p_t(c\|c_0^{t-1}, x_1^T)$	$p_t(c\|c_0^{t-1}, c_{t+1}^T, x_1^T)$

- experiments: typically $p_t(c|x_1^T)$
- exploitation of recurrency within each layer

Outline

Outline

Hybrid Approach

consider modeling the acoustic vector x_t in an HMM:

- phonetic labels (allophones, sub-phones): $(s, W) \rightarrow \alpha = \alpha_{sW}$
  (typical approach: decision trees, e.g. CART):

$$p(x_t|s, W) = p(x_t|\alpha_{sW})$$

- re-write the emission probability for label α and acoustic vector x_t:

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

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

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

- justification:
  - easier learning problem: labels $\alpha = 1, ..., 5000$ vs. vectors $x_t \in \mathbb{R}^{D=40}$
  - well-known result in pattern recognition (but ignored in ASR!)

History: Artificial Neural Networks in Acoustic Modeling

approaches in ASR:

- [Waibel & Hanazawa$^+$ 1988]: phoneme recognition using time-delay neural networks
- [Bridle 1989]: softmax operation for probability normalization in output layer
- [Bourlard & Wellekens 1990]:
  - for squared error criterion, NN outputs can be interpreted as class posterior probabilities (rediscovered: Patterson & Womack 1966)
  - they advocated the use of MLP outputs to replace the emission probabilities in HMMs
- [Robinson 1994]: recurrent neural network
  - competitive results on WSJ task
  - his work remained a singularity in ASR
- ...

experimental situation:
until 2011, NNs were never really competitive with(out) Gaussian Mixture Models

History: Artificial Neural Networks in Acoustic Modeling

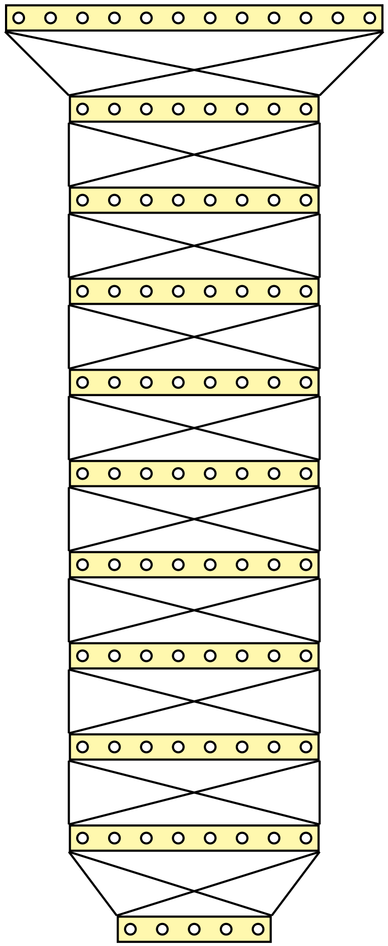
related approaches:

- [LeCun & Bengio+ 1994]: convolutional neural networks
- A. Waibel's team [Fritsch & Finke+ 1997]: hierarchical mixtures of experts
- [Hochreiter & Schmidhuber 1997]: long short-term memory neural computation (LSTM RNN) with extensions [Gers & Schraudolph+ 2002]

(second) renaissance of NN: concepts of deep learning and related ideas:

- [Hermansky & Sharma 1998]: TRAPS: learning temporal patterns of spectral energies
- [Hermansky & Ellis+ 2000]: tandem approach - multiple layers of processing by combining Gaussian model and NN for ASR
- [Utgoff & Stracuzzi 2002]: many-layered learning for symbolic processing
- [Hinton & Osindero+ 2006]: introduced what they called *deep learning (belief nets)*
- [Graves & Liwicki+ 2008]: good results for LSTM RNN on handwriting task
- Microsoft Research [Seide & Li+ 2011, Dahl & Yu+ 2012]:
  - combined Hinton's deep learning with hybrid approach
  - significant improvement by deep MLP on a large-scale task
- since 2012: other teams confirmed reductions of WER by 20% to 30%

What is Different Now after 25 Years? - A (Simplified) Summary



Comparison of today's systems vs. 1989-1994:

- number of hidden layers: 10 (or more) rather than 2-3
- number of output nodes: 5000 (or more) rather than 50
- optimization strategy: practical experience and heuristics, e.g. layer-by-layer pretraining
- computation power: much more

Terminology (for feedforward and recurrent nets):

- deep neural network
- deep learning

Outline

Training Strategies

Frame level: cross-entropy $\log p_\theta(\alpha_{s_t,W}|x_t)$

- required: single best path for each training sentence
- re-alignments during backprop learning: yes ... occasionally ... no

$\rightarrow$ simple implementation due to decoupling of best path and backprop

Sentence level: *discriminative sequence training:*

- includes language model $p(W)$
- requires sentence level posterior probability $p(W|x_1^T)$
- improvement: use exponents for language model, transition probabilities and acoustic model
- approximations: single best path, lattice with/without re-computation, ...
- three types of discriminative criteria:
  - logarithm of posterior probability
  - MPE applied to phones: 1 out of ˜50
  - MPE applied to CART labels: 1 out of ˜5000

$\rightarrow$ complex implementation

Outline

Experimental Setup

Experimental conditions:

- QUAERO task: English broadcast news and conversations (evaluation campaign 2011)
- training data: two conditions: 50 and 250 hours
- test data: dev and eval sets, each 3 hours
- language model: vocabulary size of 150k (OOV: 0.4%) and perplexity of 130

Baseline Gaussian mixture HMM based acoustic model:

- feature vector: 16 MFCC (mel frequency cepstral coefficients)
- augmented feature vector: $9 \cdot 16 = 144$
- high-performance baseline system:
  Gaussian mixtures with pooled diagonal covariance matrix:
  - reduction by LDA to 45-dimensional vector
  - 4501 CART labels
  - 680k densities
  - total number of free parameters: $680k \cdot (45 + 1) = 31.3M$

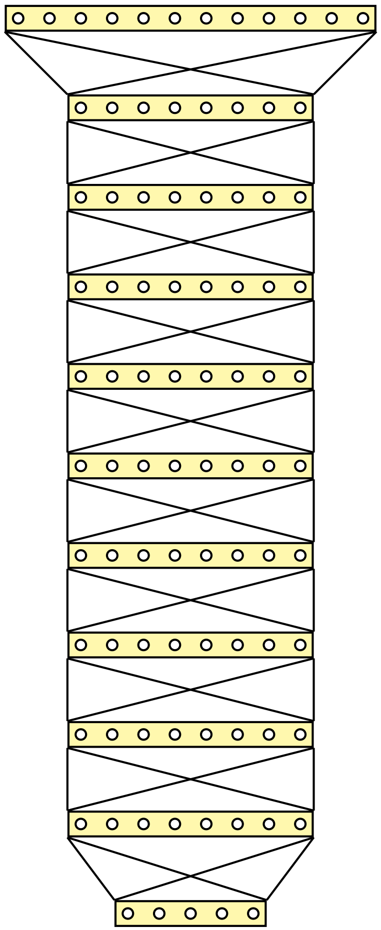
Gaussian Mixture Models (GMM): Influence of Training Criteria

Training Criterion	WER [%]			
	50h		250h	
	dev	eval	dev	eval
Maximum likelihood	24.4	31.6	22.1	28.6
MMI at frame level	23.9	30.9	22.1	28.6
MMI at sentence level	24.1	31.2	21.7	28.1
Minimum phone error	23.6	30.2	20.4	26.2

remarks:

- best improvement over maximum likelihood:
  5-10% relative by MPE (Minimum Phone Error)
- comparative evaluations in QUAERO:
  competitive results with LIMSI Paris and KIT Karlsruhe

Deep MLP: Number of Hidden Layers



- WER vs. number of hidden layers for 50-h training corpus
- Structure of MLP:
  - input dimension: 493 (window + derivatives)
  - 2000 nodes per hidden layer
  - nonlinearity: sigmoid
  - number of parameters for 6-layer MLP:

$$\begin{aligned} & 493 \cdot 2000 \\ & +5 \cdot 2000^2 \\ & +2000 \cdot 4501 \\ = & \; 30\text{M} \end{aligned}$$

- improvement over best GMM: 20% relative

hidden layers	WER [%]	
	dev	eval
1	24.5	31.3
2	22.0	28.3
3	20.5	26.7
4	19.8	26.1
5	20.1	26.0
6	19.6	25.4
7	19.7	25.5
8	19.6	25.7
9	19.3	25.3
best GMM	23.6	30.2

Discriminative Sequence Training: MPE vs. CE

Comparison of two training criteria (MLP with 6 hidden layers, 2000 nodes each):

- baseline: cross-entropy = frame MMI
- MPE: minimum phone error (context of pron. lexicon and language model)

Model	Criterion	WER [%]			
		50h		250h	
		dev	eval	dev	eval
MLP	frame MMI	19.6	25.4	15.2	20.4
	MPE	17.5	23.3	14.1	19.2
best GMM	MPE	23.6	30.2	20.4	26.4

experimental result: improvement of 5-10% by MPE over frame MMI

Activation Function: Sigmoid vs. RLU

- activation functions:
  - sigmoid function: $u \to f(u) = 1/(1 + e^{-u})$
  - RLU=rectified linear unit: $u \to f(u) = \max\{0, u\}$
- structure of MLP:
  - 6 hidden layers, each with 2000 nodes
  - training condition:
    - (frame-wise) cross-entropy
    - L2 regularization (weight decay): important
    - momentum term
- word error rates for activations functions: sigmoid vs. RLU:

activation function	WER [%]			
	50h		250h	
	dev	eval	dev	eval
sigmoid	19.6	25.4	15.2	20.4
RLU	17.7	23.5	14.7	19.6
best GMM	23.6	30.2	20.4	26.4

- experimental result: improvement of 5-10% by RLU over sigmoid

Deep LSTM-RNN

50h QUAERO training corpus:

- baseline: best MLP:
  - input: 50 Gammatone features
  - 9 hidden layers
  - RLU
  - training criterion: cross-entropy
- LSTM-RNN structure:
  - input: 50 Gammatone features
  - training criterion: cross-entropy
  - bidirectional with several hidden layers
  - 500 nodes per hidden layer
  - training on a single GPU
- eval improvements:
  - 14% relative over MLP
  - 42% relative over GMM

LSTM layers	#params	time / epoch	WER [%]	
			dev	eval
1	6.7M	0:28h	17.6	22.7
2	12.7M	1:00h	14.6	18.8
3	18.7M	1:11h	14.0	18.4
4	24.7M	1:33h	13.5	17.7
5	30.7M	1:48h	13.6	17.7
6	36.7M	2:10h	13.5	17.5
7	42.7M	2:36h	13.8	18.0
8	48.7M	3:14h	14.2	18.4
best MLP (9x2000)	42.7M	0:35h	15.3	20.3
best GMM	31.3M	–	23.6	30.2

Effect of ANNs in Acoustic Modelling

Compare three types of emission models in HMMs:

- GMM: Gaussian mixture model
- MLP: deep multi-layer perceptron
- LSTM RNN: recurrent neural network with long short-term memory

Experimental results for QUAERO English 2011:

approach	layers	WER[%]
conventional: best GMM	–	30.2
hybrid: best MLP	9	20.3
hybrid: best LSTM RNN	6	17.5

Remarks:

- comparative evaluations in QUAERO 2011:
  competitive results with LIMSI Paris and KIT Karlsruhe
- best improvement over Gaussian mixture models
  by 40% relative using an LSTM RNN

Outline

Human Language Technology: Overview & History

Statistical Approach

Neural Network and Statistical Approach

Deep Learning for Acoustic Modelling

Deep Learning for Language Modelling
- History of Neural Networks in Language Modeling
- Perplexity vs. Word Error Rate
- Neural Network based Language Modeling
- Empirical Overview of Current Methods

Current State-of-the-Art in ASR

References

Outline

History of Neural Networks in Language Modeling

- [Nakamura & Shikano 1989]:
  English word category prediction based on neural networks.
- [Castano & Vidal^{+} 1993]:
  Inference of stochastic regular languages through simple recurrent networks
- [Bengio & Ducharme^{+} 2000]:
  A neural probabilistic language model
- [Schwenk 2007]:
  Continuous space language models
- [Mikolov & Karafiat^{+} 2010]:
  Recurrent neural network based language model
- RWTH Aachen [Sundermeyer & Schlüter^{+} 2012]:
  LSTM recurrent neural networks for language modeling
- RWTH Aachen [Sundermeyer & Tüske^{+} 2014]:
  long range LM rescoring beyond N-best lists

Today: neural network based language models show competitive results.

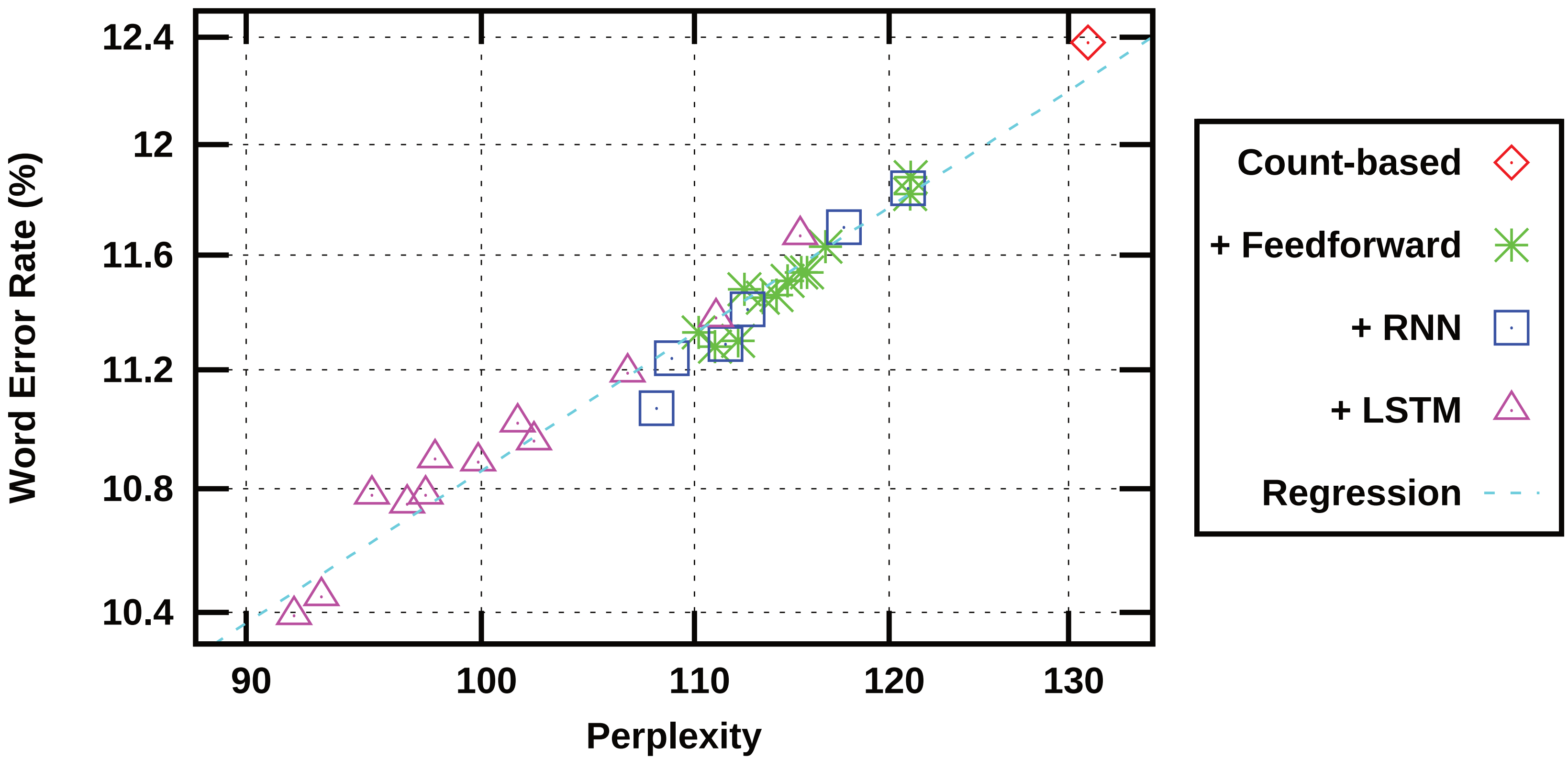
Outline

Reminder: perplexity (PP)

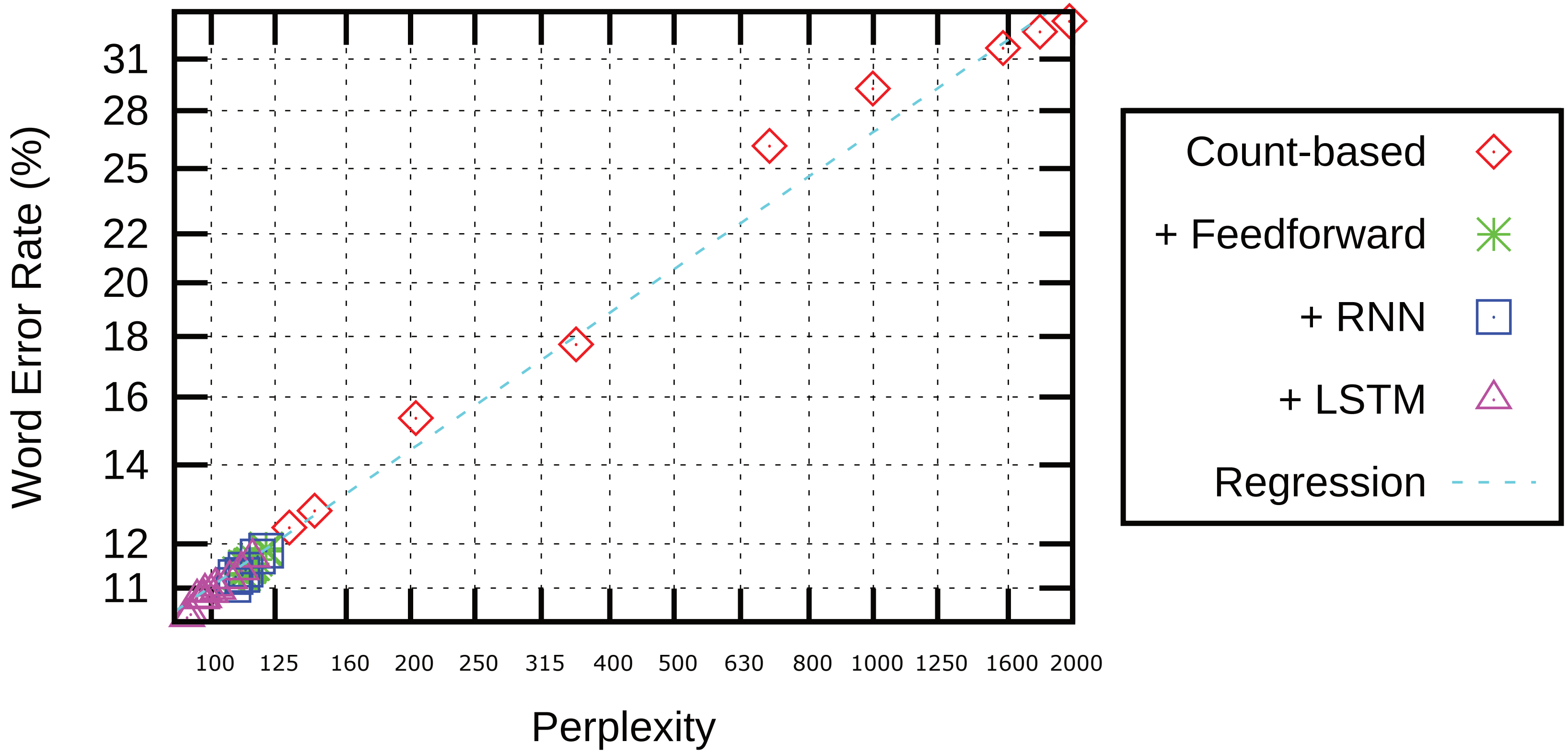
- geometric average of inverse probability → interpretation: average effective vocabulary size

$$PP := \left(p(w_1^N)\right)^{-1/N} = \left(\prod_{n=1}^{N} p(w_n|w_1^{n-1})\right)^{-1/N}$$

define w_1^0 as empty sequence

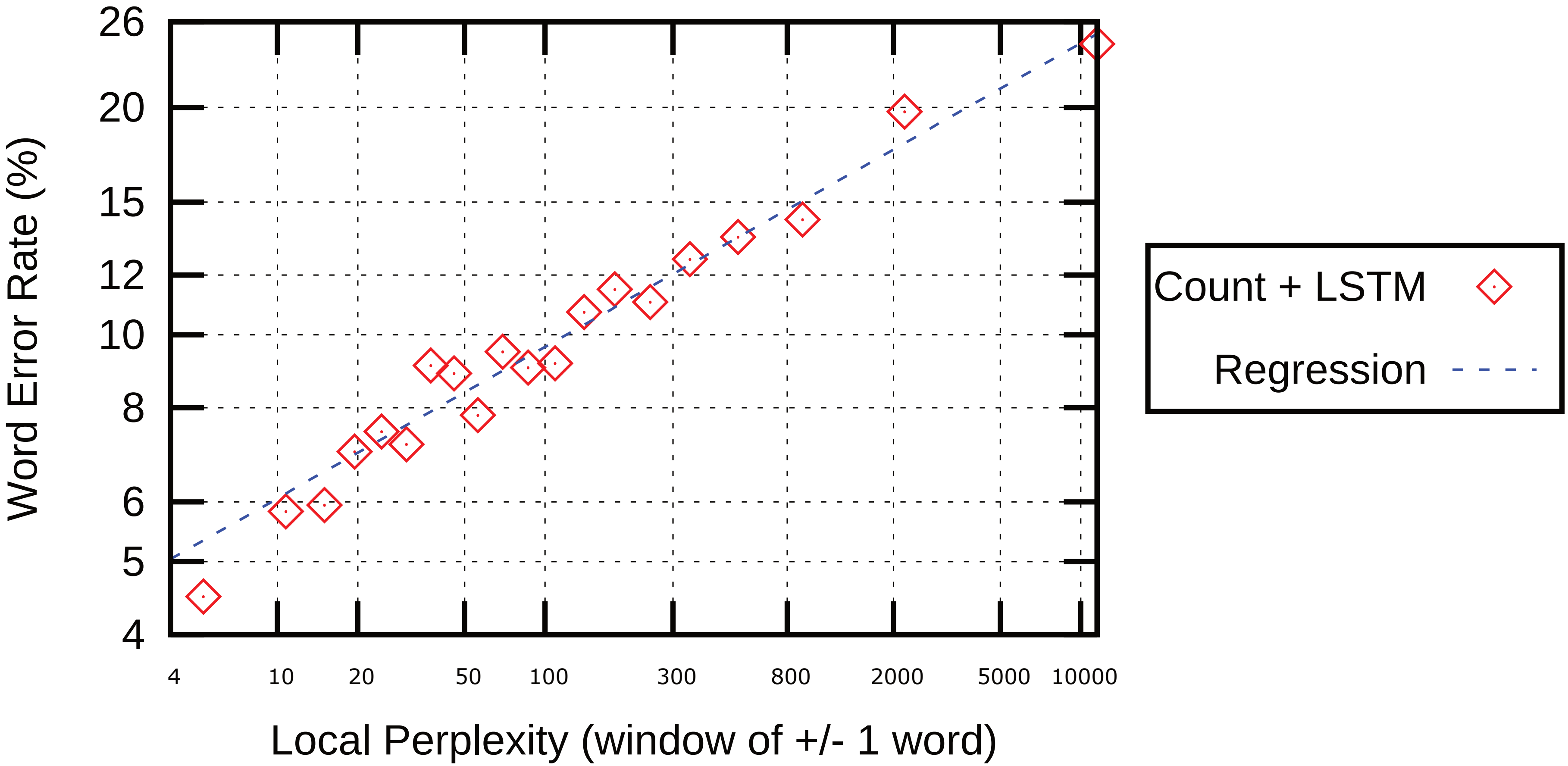


Extended Range: Perplexity vs. Word Error Rate



- empirical results, originally proposed by [Klakow & Peters 2002]
- analytical error bound exists [Schlüter & Nußbaum-Thom$^+$ 2013] (upper bound only)
- proof of approximate power law still missing

Word Error Rate vs. Local Perplexity (3-word window, 20 bins)



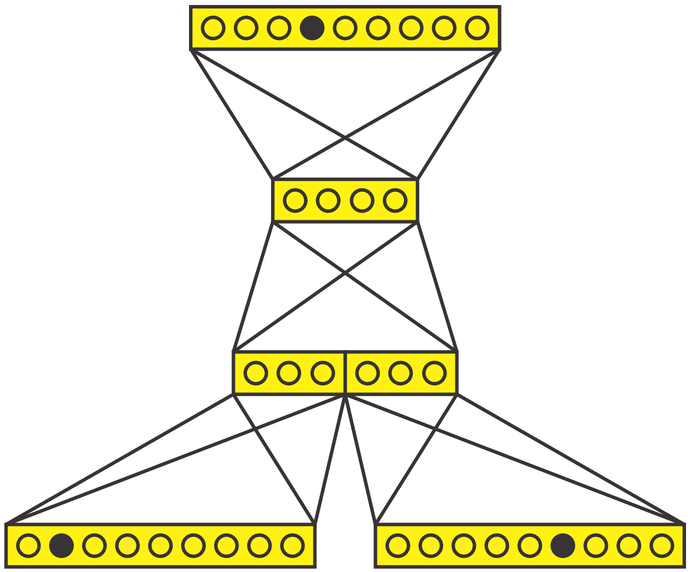
Outline

Neural Network based Language Modeling

- distinguish:
  - *sub-symbolic* processing: speech/audio, text images, image/video (computer vision)
  - *symbolic processing:* language modeling (and machine translation)
- word sequence $w_1^N := w_1 \ldots w_n \ldots w_N$
- language model: conditional probability $p(w_n|w_0^{n-1})$ (with artificial start symbol w_0):

$$p(w_1^N) = \prod_{n=1}^{N} p(w_n|w_0^{n-1})$$

- approaches to modeling $p(w_n|w_0^{n-1})$
  - count models (Markov chain):
    - limit history w_0^{n-1} to k predecessor words
    - smooth relative frequencies (e.g. SRI toolkit)
  - MLP models:
    - limit history, too
    - use predecessor words as input to MLP
  - RNN models:
    - unlimited history! [Mikolov & Karafiat+ 2010]

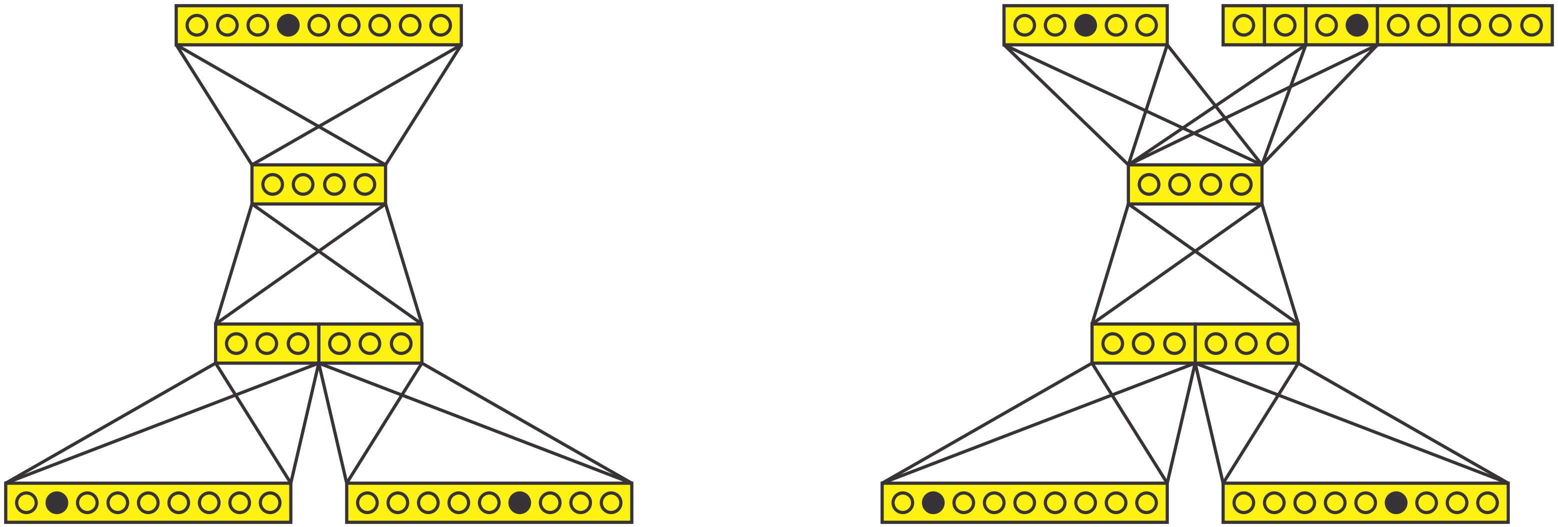


Structure of Neural Network for Language Modeling

- input layer: k predecessor words with 1-of-V coding (V = vocabulary size)
- first layer: *projection layer*
  - idea: dimension reduction (e.g. from 150k to 600!)
  - a linear operation (matrix multiplication) without sigmoid activation
  - shared accross all predecessor words of the history h
- output layer:
  - conditional probability of language model $p(w|h)$
  - softmax operation for normalization
- training criterion:
  - perplexity: equivalent to cross-entropy
  - early stopping using cross-validation on dev corpus
- properties of softmax operation:
  - computationally expensive (sum over full vocabulary)
  - remedy: word classes (automatically trained)
  - normalized outputs of softmax fit nicely into perplexity criterion

Word Classes

MLP w/o and with Word Classes: Trigram LM



factorization of conditional language model probability $p(w|h)$ for each history h:

$$p(w|h) = p(g|h) \cdot p(w|g,h)$$

using a unique word class g for each word w

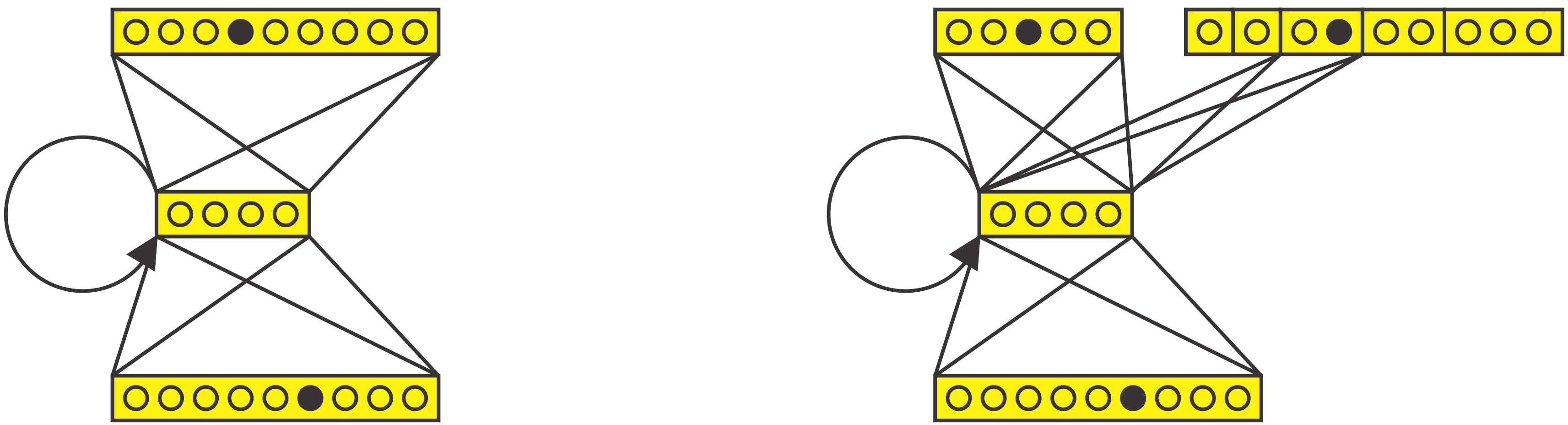
Word Classes

RNN without and with Word Classes

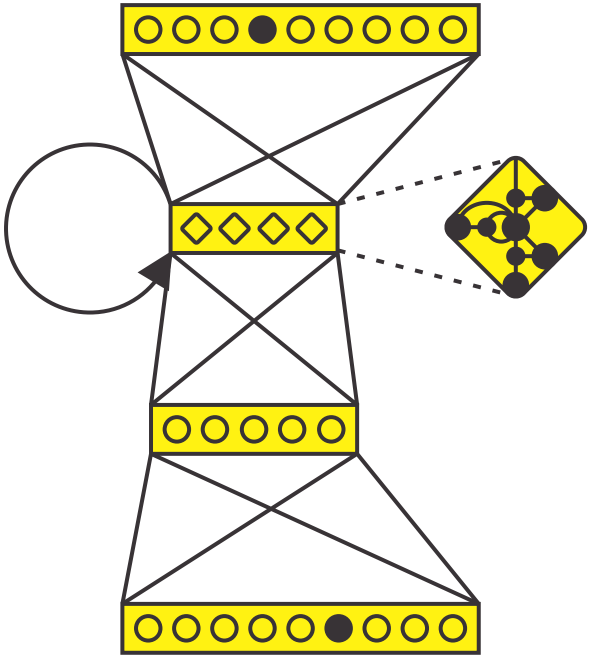
- NN with memory for sequence processing
- left-to-right processing of word sequence $w_1 \dots w_n \dots . w_N$

$$p(w_1^N) = \prod_n p(w_n | w_0^{n-1}) = \prod_n p(w_n | w_{n-1}, h_{n-1})$$

- input to RNN in position n:
  - output h_{n-1} of hidden layer at position $(n-1)$
  - immediate predecessor word w_{n-1}



LSTM RNN [Hochreiter & Schmidhuber 1997, Gers & Schraudolph+ 2002]



refinement of RNN:
LSTM = long-short term memory

- RNN: problems with vanishing/exploding gradients
- remedy: cells with gates rather than nodes
- details: see literature

Outline

Empirical Overview of Current Methods

- results on QUAERO English (like before):
  - vocabulary size: 150k words
  - training text: 50M words
  - dev and eval sets: 39k and 35k words
- MLP: structure:
  - projection layer: 300 nodes
  - hidden layer: 600 nodes
  - size of MLP is dominated by input and output layers: $150k \cdot 300 + 600 \cdot 150k = 135M$
- RNN (and LSTM RNN): structure
  - projection and hidden layer: each 600 nodes
  - size of RNN is dominated by input and output layers: $150k \cdot 600 + 600 \cdot 150k = 180M$

perplexity PPL on dev data:

approach	hidden layers	PPL
count model	–	163.7
10-gram MLP	1	136.5
	2	130.9
RNN	1	125.2
LSTM-RNN	1	107.8
	2	100.5

observation:
(huge) improvement by 40%

Complexity: Computation Times

Training times (without GPUs!) for training corpus of 50 Million words:

Models	PPL	CPU Time (Order)
Count model	163.7	30 min
MLP	136.5	1 week
LSTM-RNN	107.8	3 weeks

- problem: high computation times
- remedy: two types of language models:
  - count model: trained on a huge corpus: 3.1 Billion words
  - NN models: trained on a small corpus: 50 Million words
- resulting language model:
  linear interpolation of *two* models

Interpolated Language Models: Perplexity and WER

- linear interpolation of *two* models: count model + NN model
- perplexity and word error rate on test data:

Models	PPL	WER[%]
count model	131.2	12.4
+ 10-gram MLP	112.5	11.5
+ Recurrent NN	108.1	11.1
+ LSTM-RNN	96.7	10.8
+ 10-gram MLP with 2 layers	110.2	11.3
+ LSTM-RNN with 2 layers	92.0	10.4

- experimental result:
  - significant improvements by NN language models
  - best improvement in perplexity: 30% reduction (from 131 to 92)
  - best improvement in WER: 16% reduction (from 12.4% to 10.4%)
  - empirical observation:
    power law between WER and perplexity (cube to square root)

Outline

Outline

Overall Improvements by ANNs in ASR

QUAERO English Eval 2013

Language Model	PP	Acoustic Model	WER[%]
Count Fourgram	131.2	Gaussian Mixture	19.2
		deep MLP	10.7
		LSTM RNN	10.4
+ LSTM-RNN	92.0	Gaussian Mixture	16.5
		deep MLP	9.3
		LSTM RNN	9.3

Remarks:

- overal improvements by ANNS: 50%
- lion's share of improvement: acoustic model
- acoustic input features: optimized for model

Outline

Recent Switchboard State-of-the-Art Systems

Acoustic modeling

- convolutional models:
  - visual geometry group (VGG) - very deep convolutional network (adopted from CV)
  - residual nets (ResNet) - even deeper, incl. short-cut connections (adopted from CV)
  - layer-wise context expansion with attention (LACE) - TDNN + short-cuts + attention mask
- bidirectional long-short term memory (BLSTM) recurrent network (IBM+MSR)

Language modeling

- *N*-gram vs. LSTM-NN

Experimental results:

- challenging task
- training on 2000h
- single systems
- sites compared:
  - IBM Research [Saon & Kurata+ 17]
  - Microsoft Research (MSR) [Xiong & Droppo+ 17]

site	acoustic model	LM, WER [%] *N*-gram SWB	CH	LSTM RNN SWB	CH
IBM	BLSTM	7.2	12.7	-	-
	ResNet	7.6	14.5	-	-
MSR	BLSTM	8.3	14.9	6.7	13.0
	ResNet	8.6	14.8	6.6	12.5
	VGG	9.1	15.7	7.1	13.2
	LACE	8.4	15.0	6.7	13.0

Outline

Human - Machine Comparison

How does state-of-the-art ASR compare against human performance?

- current best ASR systems obtained using system combination
- two human speech recognition studies

Results on Switchboard task cited from

- IBM Research [Saon & Kurata+ 17]
- Microsoft Research (MSR) [Xiong & Droppo+ 17]

recognition	site	WER [%] SWB	WER [%] CH
machine	MSR	5.8	11.0
	IBM	5.5	10.3
human	MSR	5.9	11.3
	IBM	5.1	6.8

Thank you for your attention

Any questions?

Outline

References

- O. Abdel-Hamid, A.R. Mohamed, H. Jiang, G. Penn: "Applying convolutional neural networks concepts to hybrid NN-HMM model for speech recognition," *IEEE Intern. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4277–4280, Mar. 2012.
- J. Anderson: "Logistic Discrimination," *Handbook of Statistics 2*, P.R. Krishnaiah and L.N. Kanal, eds., pp. 169–191, North-Holland, 1982.
- M. Auli, M. Galley, C. Quirk, G. Zweig: "Joint Language and Translation Modeling with Recurrent Neural Networks," *Proc. Conf. on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1044-1054, Seattle, Washington, WA, Oct. 2013.
- 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. R. Bahl, P. F. Brown, P. V. de Souza, R. L. Mercer: Maximum mutual information estimation of hidden Markov parameters for speech recognition. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), Tokyo, pp.49-52, April 1986.
- Y. Bengio, R. De Mori, G. Flammia, R. Kompe: "Global optimization of a neural network - hidden markov model hybrid," *IEEE Transactions on Neural Networks*, Vol. 3, pp. 252–259, Mar. 1991.
- 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.
- Y. Bengio, P. Lamblin, D. Popovici, H. Larochelle: "Greedy layer-wise training of deep networks," *Advances in neural information processing systems*, Vol. 19: Proceedings of the 2016 conference, pp. 153–160, 2007.
- Y. Bengio, P. Simard, P. Frasconi: "Learning long-term dependencies with gradient descent is difficult," *IEEE Transactions on Neural Networks*, Vol. 5, No. 2, pp. 157–166, 1994.

References

- R. Botros, K. Irie, M. Sundermeyer, H. Ney: On Efficient Training of Word Classes and Their Application to Recurrent Neural Network Language Models. Interspeech, pp.1443-1447, Dresden, Germany, Sep. 2015.
- H. Bourlard, C. J. Wellekens: 'Links between Markov Models and Multilayer Perceptrons', in D.S. Touretzky (ed.): "Advances in Neural Information 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.

References

- P. F. Brown, S. A. Della Pietra, V. J. Della Pietra, R. L. Mercer: Mathematics of Statistical Machine Translation: Parameter Estimation. Computational Linguistics, Vol. 19.2, pp. 263-311, June 1993.
- 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.
- M. Castano, F. Casacuberta: A connectionist approach to machine translation. European Conf. on Speech Communication and Technology (Eurospeech), pp. 91–94, Rhodes, Greece, Sep. 1997.
- M. Castano, F. Casacuberta, E. Vidal: Machine translation using neural networks and finite-state models. Int. Conf. on Theoretical and Methodological Issues in Machine Translation (TMI), pp. 160-167, Santa Fe, NM, July 1997.
- W. Chan, N. Jaitly, Q. V. Le, O. Vinyals: "Listen, Attend and Spell," *arXiv preprint*, arXiv:1508.01211, Aug. 2015.

References

- X. Chen, A. Eversole, G. Li, D. Yu, F. Seide: “Pipelined Back-Propagation for Context-Dependent Deep Neural Networks,” *Interspeech*, pp. 26–29, Portland, OR, Sep. 2012.
- K. Cho, B. Gulcehre, D. Bahdanau, F. Schwenk, Y. Bengio: “Learning phrase representations using RNN encoderÂdecoder for statistical machine translation,” *Conf. on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1724–1734, Doha, Qatar, Oct. 2014,
- G. Cybenko: “Approximation by Superpositions of a Sigmoidal Function,” *Mathematics of Control, Signals and Systems*, Vol. 2, No. 4, pp. 303–314, 1989.
- 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.
- J. Dean, G. Corrado, R. Monga, K. Chen, M. Devin, M. Mao, M. Ranzato, A. Senior, P. Tucker, K. Yang, Q. Le, A. Ng: “Large Scale Distributed Deep Networks,” in F. Pereira, C. Burges, L. Bottou, K. Weinberger (eds.): *Advances*

References

in Neural Information Processing Systems (NIPS), pp. 1223–1231, Nips Foundation, http://books.nips.cc, 2012.

- J. Devlin, R. Zbib, Z. Huang, T. Lamar, R. Schwartz, J. Makhoul: Fast and Robust Neural Network Joint Models for Statistical Machine Translation. Annual Meeting of the ACL, pp. 1370–1380, Baltimore, MA,, June 2014.
- 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.
- P. Doetsch, M. Hannemann, R. SchlÃ¼ter, H. Ney: "Inverted Alignments for End-to-End Automatic Speech Recognition," submitted to *IEEE Journal on Special Topics in Signal Processing, Special Issue on End-to-End Speech and Language Processing*, April 2017.
- P. Dreuw, P. Doetsch, C. Plahl, G. Heigold, H. Ney: "Hierarchical hybrid MLP/HMM or rather MLP features for a discriminatively trained Gaussian HMM: A comparison for offline handwriting recognition," *Intern. Conf. on Image Processing*, 2011.

References

- H. Erdogan, J. Hershey, S. Watanabe, M. Mandel, J. Le Roux: “Improved MVDR Beamforming Using Single-Channel Mask Prediction Networks,” *INTERSPEECH*, pp. 1981-1985, Sep. 2016.
- S. Espana-Boquera, M. Castro-Bleda, J. Gorbe-Moya, F. Zamora-Martinez: “Improving offline handwritten text recognition with hybrid HMM/ANN models,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 33, No. 4, pp. 767–779, Apr. 2011.
- V. Fontaine, C. Ris, J.-M. Boite: Nonlinear discriminant analysis for improved speech recognition, Eurospeech, Rhodes, Greece, Sept. 1997.
- 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.
- K. Fukushima: “Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position,” *Biological Cybernetics*, Vol. 36, No. 4, pp. 193–202, April 1980.

References

- T. Gao, J. Du, L.-R. Dai, C.-H. Lee : "Joint Training of Front-End and Back-End Deep Neural Networks for Robust Speech Recognition," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4375-4379, Apr. 2015.
- F. A. Gers, J. Schmidhuber, F. Cummin: Learning to forget: Continual prediction with LSTM. Neural computation, Vol 12, No. 10, pp. 2451-2471, 2000.
- 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.
- X. Glorot, Y. Bengio: "Understanding the difficulty of training deep feedforward neural networks," *Int. Conf. on Artificial Intelligence and Statistics*, pp. 249–256, 2010.
- V. Goel, W. Byrne: "Minimum Bayes Risk Automatic Speech Recognition," *Computer Speech and Language*, Vol. 14, No. 2, pp. 115–135, April 2000.
- I. Goodfellow, Y. Bengio, A. Courville: *Deep Learning*, Book in preparation for MIT Press, `http://www.deeplearningbook.org`, 2016.

References

- J. Goodman: "Classes for fast maximum entropy training," *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Vol. 1, pp. 561–564, Salt Lake City, UT, May 2001.
- P. Golik, P. Doetsch, H. Ney: "Cross-Entropy vs. Squared Error Training: a Theoretical and Experimental Comparison," *Interspeech*, pp. 1756–1760, Lyon, France, Aug 2013.
- 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.
- 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, 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.

References

- F. Grézl, M. Karafiát, M. Janda: "Study of probabilistic and bottle-neck features in multilingual environment," *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pp. 359–364, Waikoloa, HI, Dec. 2011.
- 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.
- G. Heigold, E. McDermott, V. Vanhoucke, A. Senior, M. Bacchiani: "Asynchronous stochastic optimization for sequence training of deep neural networks," *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5587–5591, Florence, Italy, May 2014.
- G. Heigold, R. Schlüter, H. Ney, S. Wiesler: "Discriminative Training for Automatic Speech Recognition: Modeling, Criteria, Optimization, Implementation, and Performance," *IEEE Signal Processing Magazine*, Vol. 29, No. 6, pp. 58–69, Nov 2012.
- G. Heigold, S. Wiesler, M. Nussbaum, P. Lehnen, R. Schlter, H. Ney: "Discriminative HMMs, Log-Linear Models, and CRFs: What is the Difference?"

IEEE Intern. Conf. on Acoustics, Speech, and Signal Processing (ICASSP), pp. 5546–5549, Dallas, TX, March 2010.

- H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," Journal of the Acoustical Society of America, Vol. 8, No. 4, pp. 1738–1752, 1990
- 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.
- H. Hermansky, P. Fousek: "Multi-resolution RASTA filtering for TANDEM-based ASR," Interspeech, pp. 361–364, Lisbon, Portugal, Sept. 2005.
- H. Hermansky, S. Sharma: "TRAPS - classifiers of temporal patterns," *Int. Conf. on Spoken Language Processing (ICSLP)*, pp. 1003–1006, Sydney, Australia, Dec. 1998.
- J. Heymann, L. Drude, A. Chinaev, R. Haeb-Umbach: "BLSTM Supported GEV Beamformer Front-end for the 3rd CHiME Challenge," *Proc. IEEE Automatic Speech Recognition and Understanding Workshop*, pp. 444-451, Dec. 2015.

References

- T. Higuchi, N. Ito, T. Yoshioka, T. Nakatani: 'Robust MVDR Beamforming Using Time Frequency Masks for Online Offline ASR in Noise," *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5210-5214, Mar. 2016.
- 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.
- G. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, R. Salakhutdinov: "Improving neural networks by preventing co-adaptation of feature detectors," *arXiv preprint* arXiv:1207.0580, 2012.
- J. Hochreiter: *Untersuchungen zu dynamischen neuronalen Netzen*, diploma thesis, Computer Science, TU München, June 1991.
- S. Hochreiter, J. Schmidhuber: Long short-term memory. Neural Computation, Vol. 9, No. 8, pp. 1735–1780, Nov. 1997.
- T. Hori, Y. Kubo, A. Nakamura, "Real-time One-pass Decoding with Recurrent Neural Network Language Model for Speech Recognition," *Proc. Int. Conf.*

Acoustic, Speecn and Signal Processing (ICASSP), pages 6364-6368, Florence, Italy, May. 2014.

- Hornik, K., Stinchcombe, M.B., White, H.: "Multilayer Feedforward Networks Are Universal Approximators," *Neural Networks*, Vol. 2, No. 5, pp. 359–366, Jul. 1989.
- Z. Huang, G. Zweig, B. Dumoulin, "Cache Based Recurrent Neural Network Language Model Inference for First Pass Speech Recognition," *Proc. Int. Conf. Acoustic, Speecn and Signal Processing (ICASSP)*, pages 6354-6358, Florence, Italy, May. 2014.
- K. Irie, P. Golik, R. Schlüter, H. Ney. "Investigations on Byte-Level Convolutional Neural Networks for Language Modeling in Low Resource Speech Recognition," *In IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 5740-5744, New Orleans, LA, USA, Mar. 2017.
- M. Jaderberg, K. Simonyan, A. Zisserman: "Spatial Transformer Networks," *Advances in Neural Information Processing Systems*, pp. 2017–2025, 2015.

References

- F. Jelinek: *Statistical Methods for Speech Recognition*, MIT Press, Cambridge, MA, USA, 1997.
- Y. Kim, Y. Jernite, D. Sontag, A.M. Rush, "Character-Aware Neural Language Models,"" *Proc. AAAI Conf. on Artificial Intelligence*, pages 2741-2749, Phoenix, AZ, USA, Feb. 2016.
- B. Kingsbury: "Lattice-based optimization of sequence classification criteria for neural-network acoustic modeling," *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3761–3764, Taipei, Taiwan, April 2009.
- B. Kingsbury, T. Sainath, H. Soltau: "Scalable Minimum Bayes Risk Training of Deep Neural Network Acoustic Models Using Distributed Hessian-free Optimization," *Interspeech*, Portland, OR, Sep. 2012.
- D. Klakow, J. Peters: "Testing the Correlation of Word Error Rate and Perplexity," *Speech Communication*, Vol. 38, No. 4, pp. 19–28, Sept. 2002.
- R. Kneser, H. Ney: "Improved clustering techniques for class-based statistical language modelling," *Eurospeech*, Vol. 93, pp. 973–976, Berlin, Germany, Sep. 1993.

References

- P. Koehn, F. J. Och, D. Marcu: Statistical Phrase-Based Translation. HLT-NAACL 2003, pp. 48-54, Edmonton, Canada, May-June 2003.
- M. Kozielski, P. Doetsch, H. Ney: “Improvements in RWTH’s system for off-line handwriting recognition,” *12th International Conference on Document Analysis and Recognition (ICDAR)*, pp. 935–939, Buffalo, NY, Aug. 2013.
- A. Krizhevsky, I. Sutskever, G. Hinton: “Imagenet Classification with Deep Convolutional Neural Networks,” *Advances in Neural Information Processing Systems*, pp. 1097–1105, 2012.
- H.S. Le, A. Allauzen, F. Yvon: Continuous space translation models with neural networks. NAACL-HLT 2012, pp. 39-48, Montreal, QC, Canada, June 2002.
- 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.

References

- Y. LeCun, B. Boser, J. Denker, D. Henderson, R. Howard, W. Hubbard, L. Jackel: “Backpropagation Applied to Handwritten Zip Code Recognition,” *Neural Computation*, Vol. 1, No. 4, pp. 541–551, 1989.
- K. Lee, C. Park, I. Kim, N. Kim, J. Lee, “Applying GPGPU to Recurrent Neural Network Language Model Based Fast Network Search in the Real-Time LVCSR,” *Proc. Interspeech*, pages 2102-2106, Dresden, Germany, Sep. 2015.
- 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.
- L. Mangu, E. Brill, A. Stolcke: “Finding Consensus Among Words: Lattice-Based Word Error Minimization,” *Proc. European Conf. on Speech Communication and Technology (Eurospeech)*, pp. 495–498, Sept. 1999.
- V. Manohar, D. Povey, S. Khudanpur: “Semi-supervised Maximum Mutual Information Training of Deep Neural Network Acoustic Models,” Interspeech, Dresden, Germany, Sept. 2015.

References

- X. Mestre, M. A. Miguel: "On Diagonal Loading for Minimum Variance Beamformers," *IEEE Symposium on Signal Processing and Information Technology (ISSPIT)*, pp. 459-462, Aug. 2003.
- 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.
- A. Mohamed, G. Dahl, G. Hinton: "Acoustic Modeling Using Deep Belief Networks," *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 20, No. 1, pp. 14–22, Jan. 2012.
- V. Nair, G. Hinton: "Rectified Linear Units Improve Restricted Boltzmann Machines," *Intern. Conf. on Machine Learning (ICML)*, pp. 807–814, Haifa, Israel, June 2010.
- M. Nakamura, K. Shikano: A Study of English Word Category Prediction Based on Neural Networks. ICASSP 89, p. 731-734, Glasgow, UK, May 1989.

References

- H. Ney, U. Essen, R. Kneser: “On the estimation ofsmall probabilities by leaving-one-out,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 17, No. 12, pp. 1202–1212, 1995.
- H. Ney: “On the Relationship between Classification Error Bounds and Training Criteria in Statistical Pattern Recognition,” *Proc. Iberian Conference on Pattern Recognition and Image Analysis (IbPRIA)*, pp. 636–645, Puerto de Andratx, Spain, June 2003.
- F. J. Och, H. Ney: A Systematic Comparison of Various Alignment Models. *Computational Linguistics,* Vol. 29, No. 1, pp. 19-51, March 2003.
- F. J. Och, H. Ney: The Alignment Template Approach to Statistical Machine Translation. *Computational Linguistics*, Vol. 30, No. 4, pp. 417-449, Dec. 2004.
- F. J. Och, C. Tillmann, H. Ney: Improved Alignment Models for Statistical Machine Translation. Joint ACL/SIGDAT Conf. on Empirical Methods in Natural Language Processing and Very Large Corpora, College Park, MD, pp. 20-28, June 1999.

References

- J.J. Odell: "The Use of Context in Large Vocabulary Speech Recognition," *Ph.D. Thesis, University of Cambridge*, Mar. 1995.
- M. Oerder, H. Ney: "Word graphs: an efficient interface between continuous-speech recognition and language understanding," *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Vol. 2, pp. 119–122, Minneapolis, MN, USA, 1993.
- 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.
- M. Paulik: "Lattice-based training of bottleneck feature extraction neural networks," *Interspeech*, 2013.
- D. Povey, P. Woodland: "Minimum phone error and I- smoothing for improved discriminative training," *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Vol. 1, pp. 105–108, Orlando, FL, May 2002.

References

- 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.
- T. Robinson, M. Hochberg, S. Renals: "IPA: Improved Phone Modelling with Recurrent Neural Networks," *IEEE Intern. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Vol. I, pp. 37–40, Adelaide, Australia, Apr. 1994.
- D. Rumelhart, G. Hinton, R. Williams: "Learning Representations By Back-Propagating Errors," Nature Vol. 323, pp. 533–536, Oct. 1986.
- T. Sainath, B. Kingsbury, V. Sindhwani, E. Arisoy, B. Ramabhadran: "Low-rank matrix factorization for deep neural network training with high-dimensional output targets," *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013.
- 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, Dec. 2015.

References

- 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*, Vol. 1703.02136, 2017.
- S. Scanzio, P. Laface, L. Fissore, R. Gemello, F. Mana: “On the Use of a Multilingual Neural Network Front-End,” *Interspeech*, pp. 2711–2714, Brisbane, Australia, Sept. 2008.
- R. Schlüter, M. Nußbaum-Thom, E. Beck, T. Alkhouli, H. Ney: “Novel Tight Classification Error Bounds under Mismatch Conditions based on f-Divergence,” *Proc. IEEE Information Theory Workshop*, pp. 432–436, Sevilla, Spain, Sept. 2013.
- 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.
- R. Schlüter, I. Bezrukov, H. Wagner, H. Ney: “Gammatone Features and Feature Combination for Large Vocabulary Speech Recognition,” *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 649–652, Honolulu, HI, April 2007.

References

- T. Schultz, A. Waibel: "Language-independent and language-adaptive acoustic modeling for speech recognition," *Speech Communication*, Vol. 35, No. 1-2, pp. 31–51, Aug. 2001.
- H. Schwenk: Continuous space language models. Computer Speech and Language, Vol. 21, No. 3, pp. 492–518, July 2007.
- H. Schwenk: Continuous Space Translation Models for Phrase-Based Statistical Machine Translation. 24th Int. Conf. on Computational Linguistics (COLING), Mumbai, India, pp. 1071–1080, Dec. 2012.
- H. Schwenk , M. R. Costa-jussa, J. A. R. Fonollosa: Smooth bilingual n-gram translation. Joint Conf. on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL), pp. 430–438, Prague, June 2007.
- H. Schwenk, D. Déchelotte, J. L. Gauvain: Continuous Space Language Models for Statistical Machine Translation. COLING/ACL 2006, pp. 723–730, Sydney, Australia July 2006.

References

- F. Seide, G. Li, D. Yu: "Conversational Speech Transcription using Context-Dependent Deep Neural Networks," *Interspeech*, pp. 437–440, Florence, Italy, Aug. 2011.
- K. Simonyan, A. Zisserman: "Very Deep Convolutional Networks for Large-Scale Image Recognition," CoRR, abs/1409.1556, http://arxiv.org/abs/1409.1556, Oct. 2014.
- 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.
- A. Stolcke, Y. König, M. Weintraub: "Explicit Word Error Rate Minimization in N-Best List Rescoring," *Proc. European Conf. on Speech Communication and Technology (Eurospeech)*, pp. 163–166, Rhodes, Greece, Sept. 1997.
- H. Su, G. Li, D. Yu, F. Seide: "Error Back Propagation For Sequence Training Of Context-Dependent Deep Networks For Conversational Speech

References

Transcription," *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Vancouver, Canada, May 2013.

- M. Sundermeyer, T. Alkhouli, J. Wuebker, H. Ney: Translation Modeling with Bidirectional Recurrent Neural Networks. Conf. on Empirical Methods in Natural Language Processing (EMNLP), pp. 14–25, Doha, Qatar, Oct. 2014.
- M. Sundermeyer, H. Ney, R. Schlüter, "From Feedforward to Recurrent LSTM Neural Networks for Language Modeling," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol. 23, No. 3, pp. 13–25, March 2015.
- M. Sundermeyer, R. Schlüter, H. Ney: LSTM neural networks for language modeling. Interspeech, pp. 194–197, Portland, OR, Sep. 2012.
- M. Sundermeyer, R. Schlüter, H. Ney: "Lattice decoding and rescoring with long-span neural network language models," *Proc. Interspeech*, pages 661-665, Singapore, Sep. 2014.
- 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.

References

- I. Sutskever, O. Vinyals, Q. V. Le: "Sequence to Sequence Learning with Neural Networks," *arXiv preprint*, arXiv:1409.3215, Sep. 2014.
- 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.
- 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, 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.
- 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.

References

- Z. Tüske, M. 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, April 2015.
- Z. Tüske, R. Schlüter, H. Ney: "Multilingual Hierarchical MRASTA Features for ASR," Interspeech, pp. 2222–2226. Lyon, France, Aug. 2013.
- P. E. Utgoff, D. J. Stracuzzi: Many-layered learning. Neural Computation, Vol. 14, No. 10, pp. 2497-2539, Oct. 2002.
- F. Valente, H. Hermansky, "Hierarchical and parallel processing of modulation spectrum for ASR applications," *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 4165–4168, Las Vegas, NV, Mar./Apr. 2008.
- F. Valente, J. Vepa, C. Plahl, C. Gollan, H. Hermansky, R. Schlüter: "Hierarchical Neural Networks Feature Extraction for LVCSR System," Interspeech, pp. 42–45, Antwerp, Belgium, Aug. 2007.

References

- A. Vaswani, Y. Zhao, V. Fossum, D. Chiang: Decoding with Large-Scale Neural Language Models Improves Translation. Conf. on Empirical Methods in Natural Language Processing (EMNLP, pp. 1387–1392, Seattle, Washington, Oct. 2013.
- K. Veselý, A. Ghoshal, L. Burget, D. Povey: "Sequence-discriminative training of deep neural networks," *Interspeech*, pp. 2345–2349, Lyon, France, Aug. 2013.
- K. Veselý, M. Karafiát, F. Grézl: "Convolutive bottleneck network features for LVCSR," *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pp. 42–47, Waikoloa, HI, Dec. 2011.
- S. Vogel, H. Ney, C. Tillmann: HMM-based word alignment in statistical translation. Int. Conf. on Computational Linguistics (COLING), pp. 836-841, Copenhagen, Denmark, Aug. 1996.
- 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

- A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, K. Lang: "Phoneme Recognition: Neural Networks vs. Hidden Markov Models," *IEEE Intern. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Vol. 1, pp. 107–110, Glasgow, Scotland, April 1989.
- E. Warsitz, R. Haeb-Umach: "Blind Acoustic Beamformting Based on Generalized Eigenvalue Decomposition," *IEEE Transactions on audio, speech, and language processing*, Vol. 15, pp. 1529-1539, Jun. 2007.
- F. Weng, A. Stolcke, A. Sankar: "Efficient lattice representation and generation," *Proc. Int. Conf. on Spoken Language Processing (ICSLP)*, pages 2531-2534, Sydney, Australia, Dec. 1998.
- F. Wessel, R. Schlüter, H. Ney: "Explicit Word Error Minimization using Word Hypothesis Posterior Probabilities," *Proc. IEEE Intern. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 33–36, Salt Lake City, UT, May 2001.
- S. Wiesler, A. Richard, R. Schlüter, H. Ney: "Mean-normalized Stochastic Gradient for Large-Scale Deep Learning," *IEEE Intern. Conf. on Acoustics,*

Speech, and Signal Processing (ICASSP), pp. 180–184, Florence, Italy, May 2014.

- W. Xiong, J. Droppo, X. Huang, F. Seide, M. Seltzer, A. Stolcke, D. Yu, G. Zweig: Achieving Human Parity in Conversational Speech Recognition. *arXiv*, Vol. 1610.05256v2, Feb. 2017.
- Y. Xu, J. Du, L.-R. Dai, C.-H. Lee : "A Regression Approach to Speech Enhancement Based on Deep Neural Networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, pp. 7-19, Jan. 2015.
- D. Yu, K. Yao, H. Su, G. Li, F. Seide: "KL-Divergence Regularized Deep Neural Network Adaptation for Improved Large Vocabulary Speech Recognition," *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7893–7897, Vancouver, Canada, May 2013.
- R. Zens, F. J. Och, H. Ney: Phrase-Based Statistical Machine Translation. 25th Annual German Conf. on AI, pp. 18–32, LNAI, Springer 2002.