Broadening the Perspective of Automatic Speech Recognition using Neural Networks

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 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 Networs (**new!**)
  - ...
  - cf. `hltpr.rwth-aachen.de/web/Software`

Outline

Outline

Outline

Input Embedding for NN Language Models

Discussion:

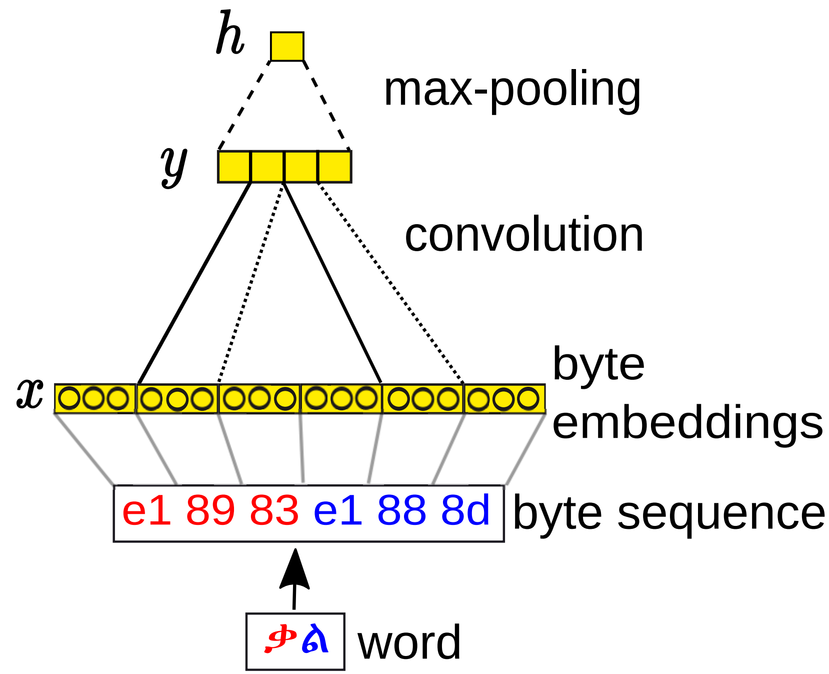
- standard: 1-of-V encoding and linear projection for each word
- problem: does not generalize to unseen words
- resort: character-level word embedding
- however: need to handle international character encodings for new languages
- idea: byte-level word embedding

Approach:

- convolution filters operate on byte level
- max-pooling to generate word-level embedding

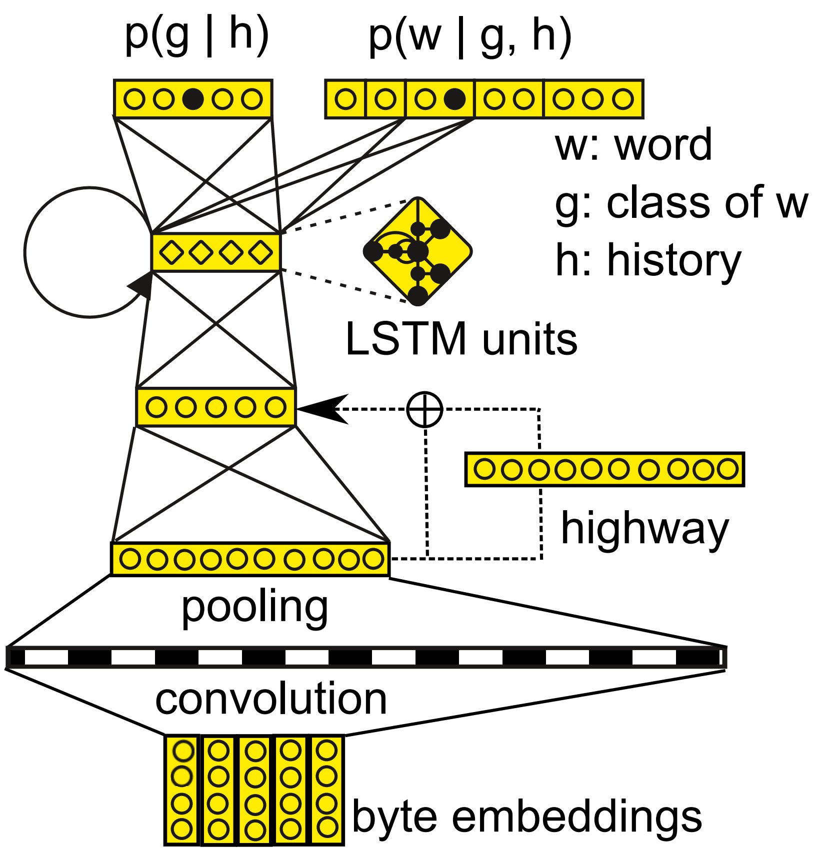
Advantage:

- no special handling for new languages



Byte-Level Convolution-based Word Embedding

Application for language modeling



- Character-aware neural LM architecture by [Kim & Jernite+ 2016].
- Classic LSTM LM with class factorized output.
- Prediction is still at **word level**.
- Standard word embedding input layer is replaced by a CNN on byte level.
- Optionally followed by a highway (adaptive interpolation) layer.

Evaluation of Byte-Level Word Embedding for ASR/Keyword Search

Experimental results from [Irie & Golik+ 2017] (Babel datasets)

word-level LM topology	Perplexity	
	Igbo	Dholuo
Baseline LSTM	103.4	144.8
+ CNN (byte)	**94.8**	136.9
+ Highway	95.9	**135.8**

ASR performance.

ID	Language	WER [%]		
		2gr	+LSTM	+CNN
306	Igbo	56.8	56.0	**55.9**
403	Dholuo	38.1	37.0	**36.9**

Keyword search performance.

ID	Language	MTWV		
		2gr	+LSTM	+CNN
306	Igbo	0.3759	0.3733	**0.3801**
403	Dholuo	0.6228	0.6245	**0.6253**

Outline

Log-Linear Interpolation of Multi-Domain Neural Network LM

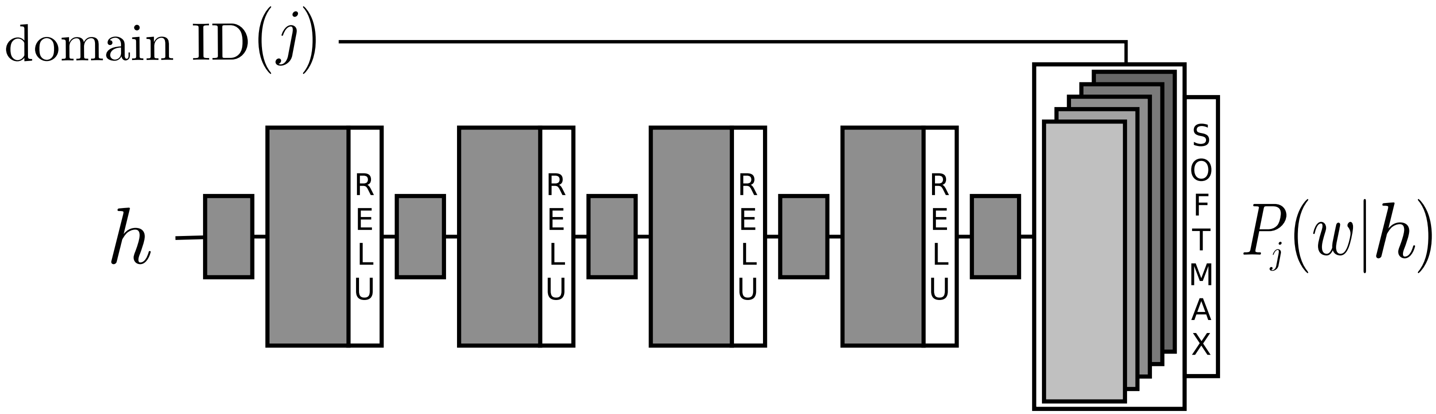
[Tüske & Irie+ 2016]

- Usual approach: linear interpolation of count LMs trained on different domains/data sets.
  - Interpolation weights optimized on target domain validation set.
  - Optimized using expectation maximization (EM) algorithm.
  - Count models are suited to be linearly combined into one single model (with union of n-grams and recomputing back-off weights)

- Goal: combination approach for neural network LMs.
  - Aiming at **single model** after interpolation of neural network LMs.
  - Linear interpolation not straightforward for NN LMs to obtain single model.
    Log-Linear combination fits better;

- Initial investigation using feed-forward NN LMs.

Joint Model



- Multiple posterior estimates
  - Active output: selected by the domain of the input vector
  - Hidden layers are shared between the domains
  - Shared vocabulary, common softmax

- Log-linear combination to obtain single overall neural network LM:
  - Leads to weighted sum of domain specific output layers.
  - Weighted sum of softmax outputs can rewritten as a single softmax output layer.

Experimental Results: Perplexities

- Training corpus: 3B words, 11 domains (Gigaword, BN/BC, TED, IWSLT, ...)
  - 50M and 2M best matching subset selected for fine-tuning
- KN 4-gram: 132.7 PPL after interpolation
- 50M LSTM-RNN: 100.5

- Retraining only multi-domain output (**log-linear!**) on the best BN, and interpolation: PPL **92.0**

LM	multi domain	log-lin. interp.	fine-tuning		PPL
			50M	2M	
50M					110.5
				×	109.0
3B					129.0
			×	×	96.2
	×				133.1
	×		×	×	95.7*
	×	×			117.6
	×	×	×	×	**94.3**

*using the best matching output

Experimental Results: WER

- Lattice generation with count model
- Lattice rescoring using `rwthlm` [Sundermeyer & Alkhouli+ 2014]
  - Traceback lattice approximation
  - Linear-interpolation of NN LM and count LM (KN 4-gram)
- Measuring word error rate
  - Acoustic model: 12-layer multilingual BN (800h), fine tuned on 250h BN/BC target data
  - Standard Viterbi (Vi.) and confusion network (CN) decoding of the lattices

Language Model	Dev			Eval		
	PPL	Vi.	CN	PPL	Vi.	CN
KN4	132.7	12.6	12.3	133.4	15.4	15.0
+ 50M FFNN	96.5	11.4	11.1	95.0	14.2	13.8
+ 3B, fine-tune	89.6	10.9	10.7	88.0	13.7	13.4
+ Multi-domain,log-lin,fine-tune	88.5	10.8	9.1	87.0	13.7	13.5
+ 50M LSTM	91.6	10.9	9.0	91.0	13.7	13.5

Outline

Background: Search Space Representation and Rescoring

Problem:

- RNN LMs imply unlimited symbol context dependency
- search space size rises exponentially with sequence length
- search space reduction requires approximation

Word graphs:

- efficient search space representation [Oerder & Ney 1993]
- enables efficient rescoring with higher-order LMs [Odell 1995]
- N-gram language models: recombination and beam-search
- unlimited context: word graph expands into (large) prefix tree
  $\rightarrow$ further approximation needed

Approach:

- pruning/approximations can be introduced to reduce the complexity.
- goal: breadth-first search - early pruning

Extension of Push-Forward Algorithm

Starting point:

- push forward algorithm from machine translation [Auli & Galley$^+$ 2013]

Approach: extract paths with RNN LM scores from the word graph

- process word graph nodes in topological order
- only retain last k words in context (k-gram context recombination)
- cardinality pruning to limit number of partial hypotheses per node
- surviving hypothesis expansion: computation of RNN state vectors ('pushing' outgoing arcs' word labels into RNN)

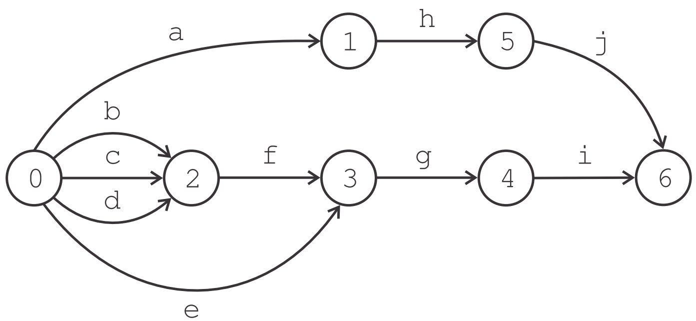
Extensions for ASR:

- integration of ASR pruning strategies (time synchronous beam pruning & look-ahead)
- processing in topological *and* temporal order
- pruning at every new time frame
- storage of rescored/expanded word graph
- presented in [Sundermeyer & Schlüter$^+$ 2014]:

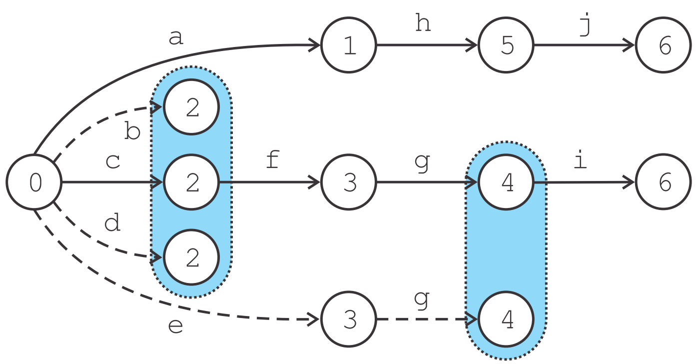
Word Graph Rescoring with RNN Language Model

Illustration from [Sundermeyer & Ney^{+} 2015]

X



Y



- X is an example lattice.
- Y is an example traceback tree when X is rescored by a RNN LM

- Pruning is illustrated by dashed lines:
  - Paths 'b-f-g-i' and 'd-f-g-i' are pruned at node 2 (middle row)
  - Path 'e-g-i' is pruned at node 4 (last row)

RNN LM Rescoring Results

Experimental results (Quaero French Test dataset) [Sundermeyer & Ney^{+} 2015]

rescoring method	WER [%]
baseline 4-gram Kneser-Ney	16.4
100-best	14.8
1000-best	14.7
word graph Rescoring (push forward)	14.6
+ Viterbi after LM scale tuning	14.5
+ confusion network decoding	**14.2**

Remarks: one-pass decoding with RNN LM?

- previous work [Huang & Zweig^{+} 2014, Hori & Kubo^{+} 2014, Lee & Park $^{+}$ 2015]
- results: WER of first pass decoding marginally better (or worse) than rescoring

Outline

Outline

Comparison and Interpretation [Tüske & Tahir+ 2015]

- State-of-the-art acoustic models (AM) are
  - Tandem acoustic models
    * Gaussian Mixture Models (GMM) are trained on the output of a neural network based features
    * Probabilistic or bottleneck (BN) tandem approach [Fontaine & Ris+ 1997, Hermansky & Ellis+ 2000, Grézl & Karafiát+ 2007]
    * **Joint training**, e.g. in [Paulik 2013]
  - Hybrid models
    * Proposed in the early 90's [Bourlard+Morgan:1993]
    * Estimates state posterior probabilities $p(s|x)$ directly
    * BN layer to train efficiently on huge number of states [Sainath & Kingsbury+ 2013]
- After careful optimization both show similar performance
- **Goal**: convert tandem into hybrid neural network representation [Tüske & Tahir+ 2015]
- **Idea**: rewrite GMM to equivalent log-linear model [Anderson 1982, Heigold & Wiesler+ 2010]
  $\rightarrow$ **softmax NN layer**

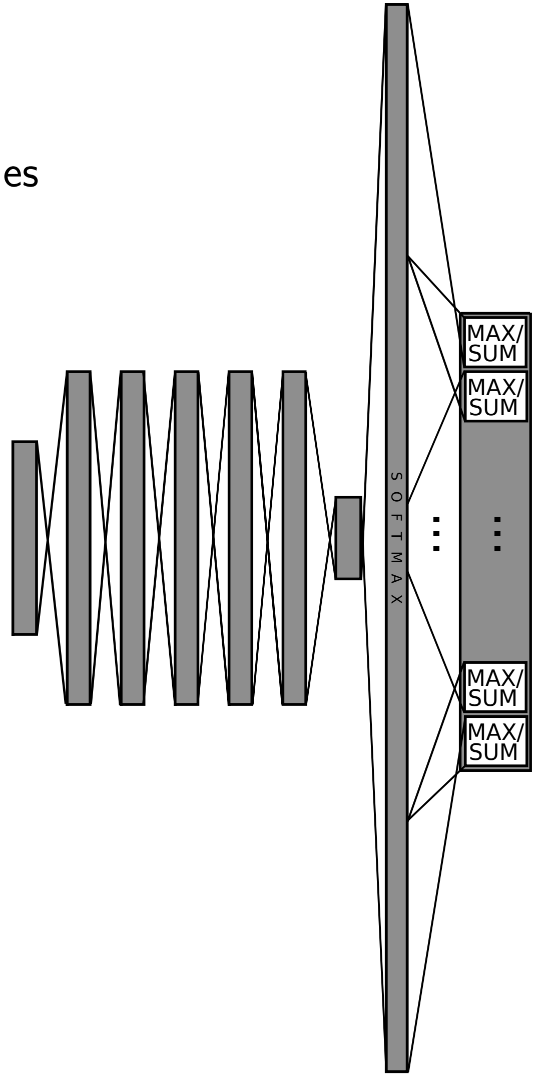
Outline

Integration of GMM into and Bottleneck DNN

- GMM with pooled covariance is a softmax layer with hidden variables
- Maximum approximation, for fast score calculation:

$$\frac{\sum_i \exp(w_{si}^T y + b_{si})}{Z(y)} \approx \left.\frac{\exp(w_{s\hat{\imath}}^T y + b_{s\hat{\imath}})}{Z(y)}\right|_{\hat{\imath}=\underset{i}{\operatorname{argmax}}(w_{si}^T y + b_{si})}$$

- No need for special element to implement:
  - sum- or max-pooling
- Efficient softmax is crucial (low-rank factorization; GPU)
  - GMM of 4500 states after 8 splits: $\sim$ 1 million nodes
- Joint training of BN and GMM:
  - Maximum likelihood training of GMM on BN features
  - Convert to LMM
  - Start the joint training
- Remark: maximum approximation with given labeling (s,i) same as classical hybrid, E-M style training is also possible



Outline

ASR Experiments

- Task: Quaero English (250h BC/BN)
- MLP structure:
  - 12 hidden layers
  - 50 dimensional Gammatone input

System	low rank	joint training	#output	#param.	split	criterion	WER [%]	
							dev	eval
Hybrid	no	–	4.5k	54.7M	-	CE	13.3	18.1
	yes			49.0M			13.5	18.2
			12.0k	52.8M			13.0	17.7
BN tandem	–	no	4.5k	613.0M	8	ML	14.2	19.0
		yes		83.5M	4	CE	13.1	17.8

- Same results with less tied-triphone states
- Smaller lexical prefix-tree

Outline

Outline

Previous Work

[Golik & Tüske+ 2015]

- large effort went into **feature engineering** for DNNs (e.g. [Seide & Li+ 2011, Yu & Yao+ 2013], ...)
- previous work [Tüske & Golik+ 2014] showed:
  - a simple fully connected 12-hidden-layers DNN performs well even **without any feature extraction**
  - WER: 22.1% (MFCC) vs. 25.5% (raw time signal)
  - first layer weights learned impulse responses of band pass filters
  - the learned filter bank roughly resembles manually defined filter bank
- **convolutional neural network (CNN)** is a natural tool that combines learning a filter bank and acoustic modeling
- research questions:
  - how much do CNNs reduce the performance gap to hand-crafted features?
  - how can we interpret the learned weights?

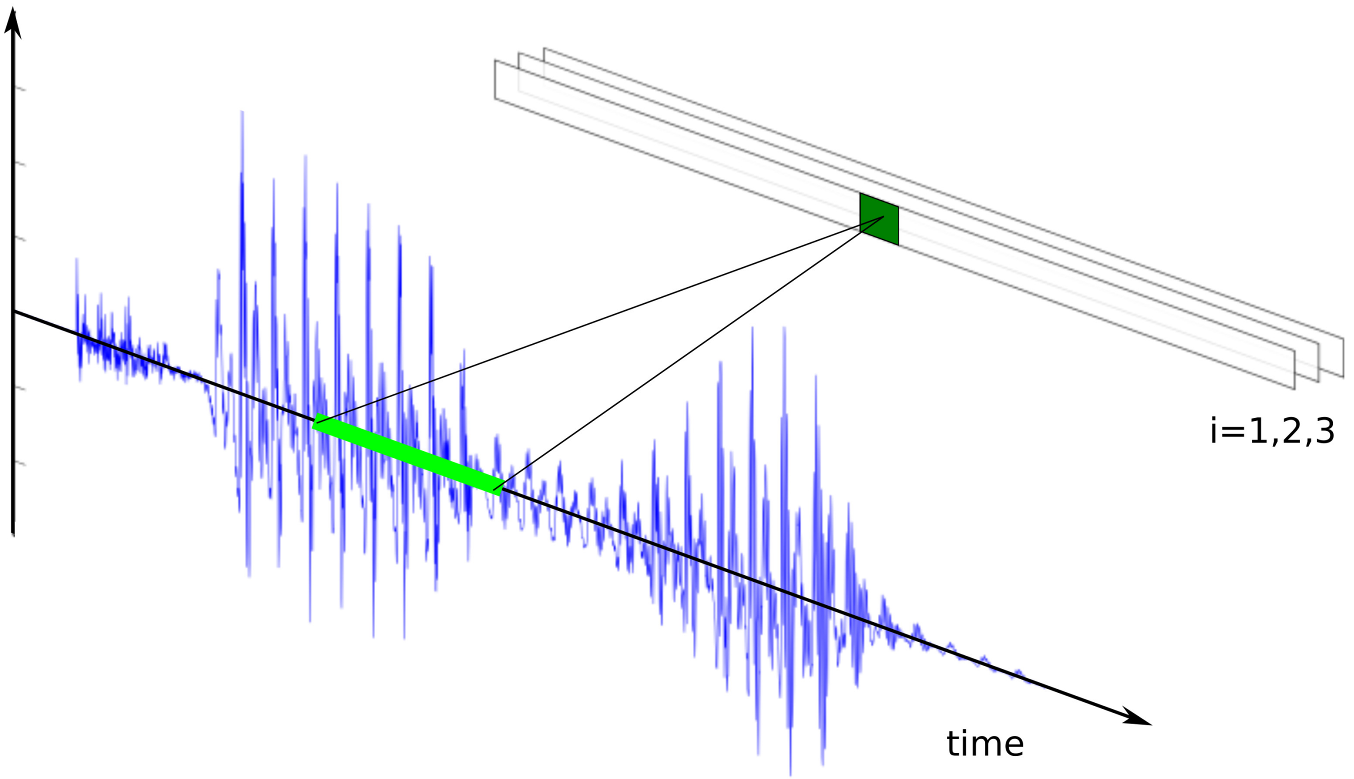
Convolutional neural networks

- CNNs were introduced to HWR about 25 years ago [LeCun & Boser+ 1989]
- today: state-of-the-art in computer vision
  ([Krizhevsky & Sutskever+ 2012, Jaderberg & Simonyan+ 2015])
- applied to speech recognition tasks by [Abdel-Hamid & Mohamed+ 2012]:
  2D filters perform convolution on a "spectrogram"
- convolution on raw time signal: 1D operation along time axis only
- output of convolutional unit i at position m:

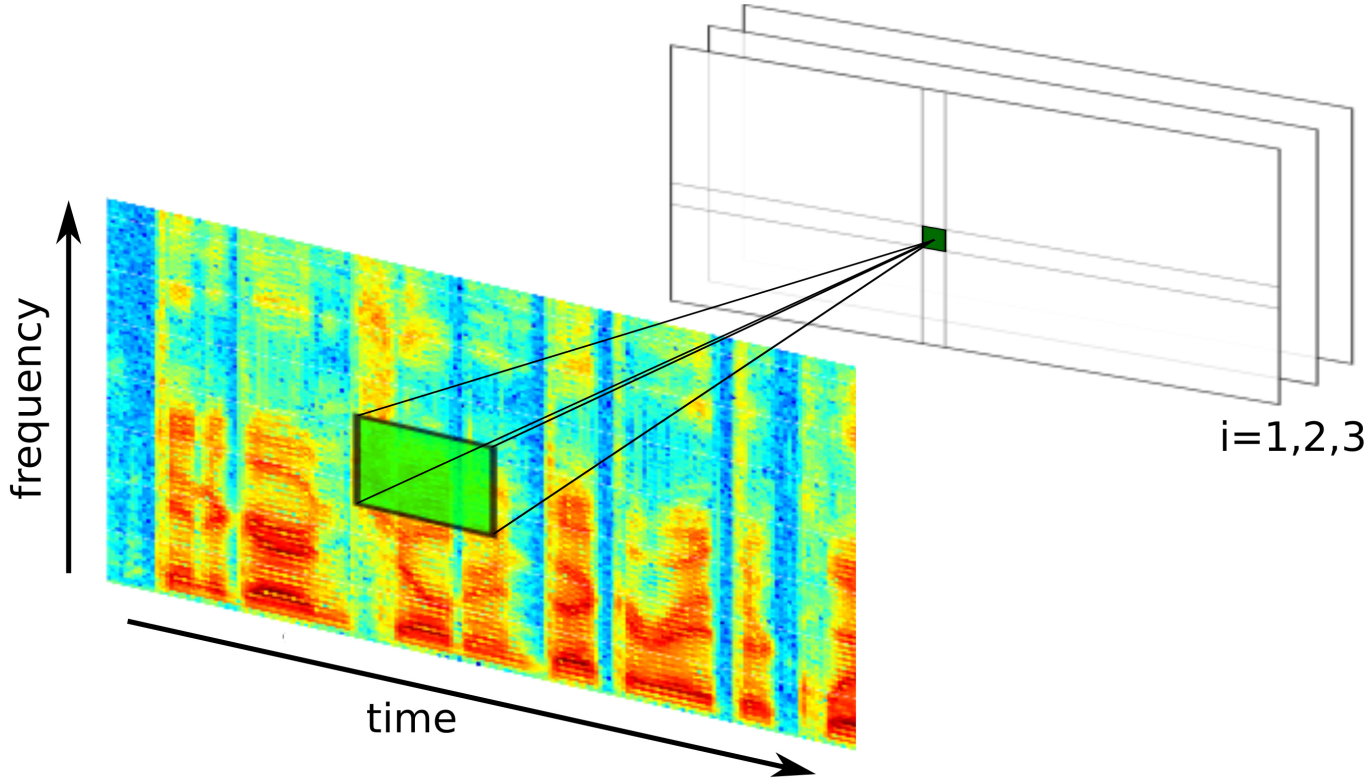
$$y_{i,m} = \sigma\left(\sum_{j=m}^{m+k-1} w_{i,j-m} x_j + b_i\right)$$

  - x_j are the PCM samples
  - $\{w_{i,\cdot}, b_i\}$: trainable parameters shared across all positions in the input
  - k is the length of the impulse response of a filter
- temporal sub-sampling by shifting m in steps of 32 and max pooling

1D convolution in time only

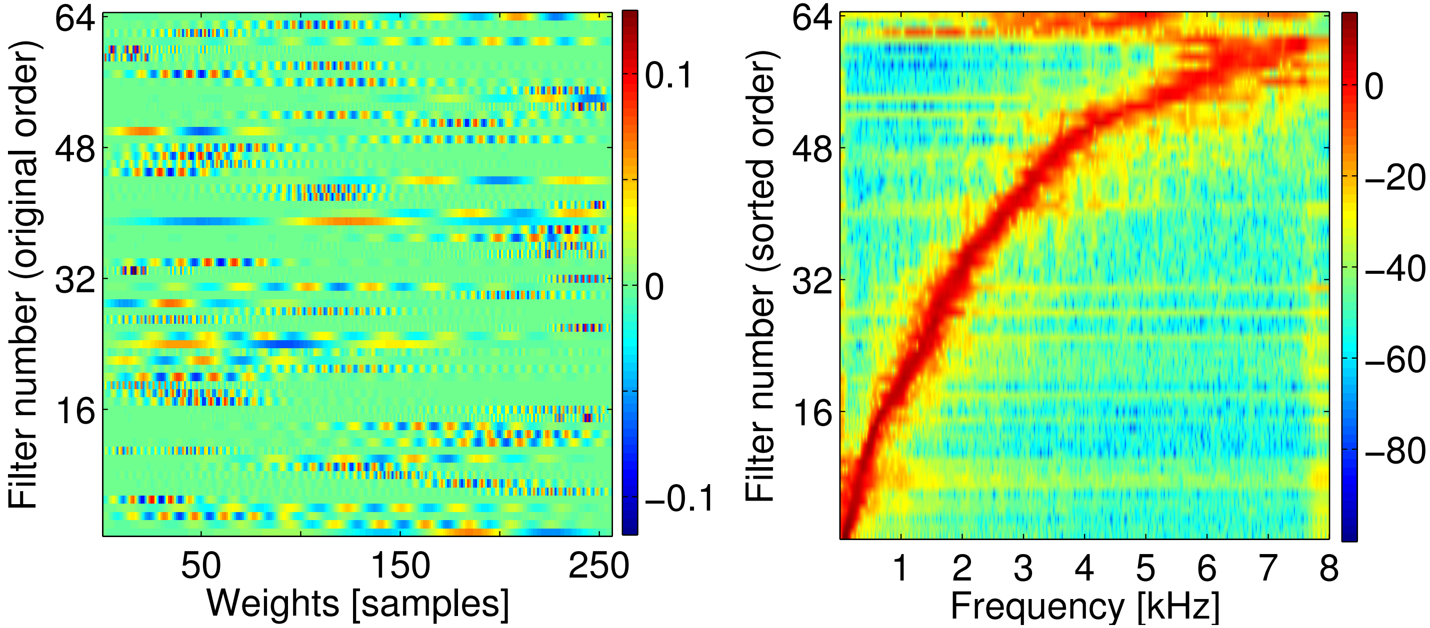


2D convolution in time/frequency (for ASR)



Outline

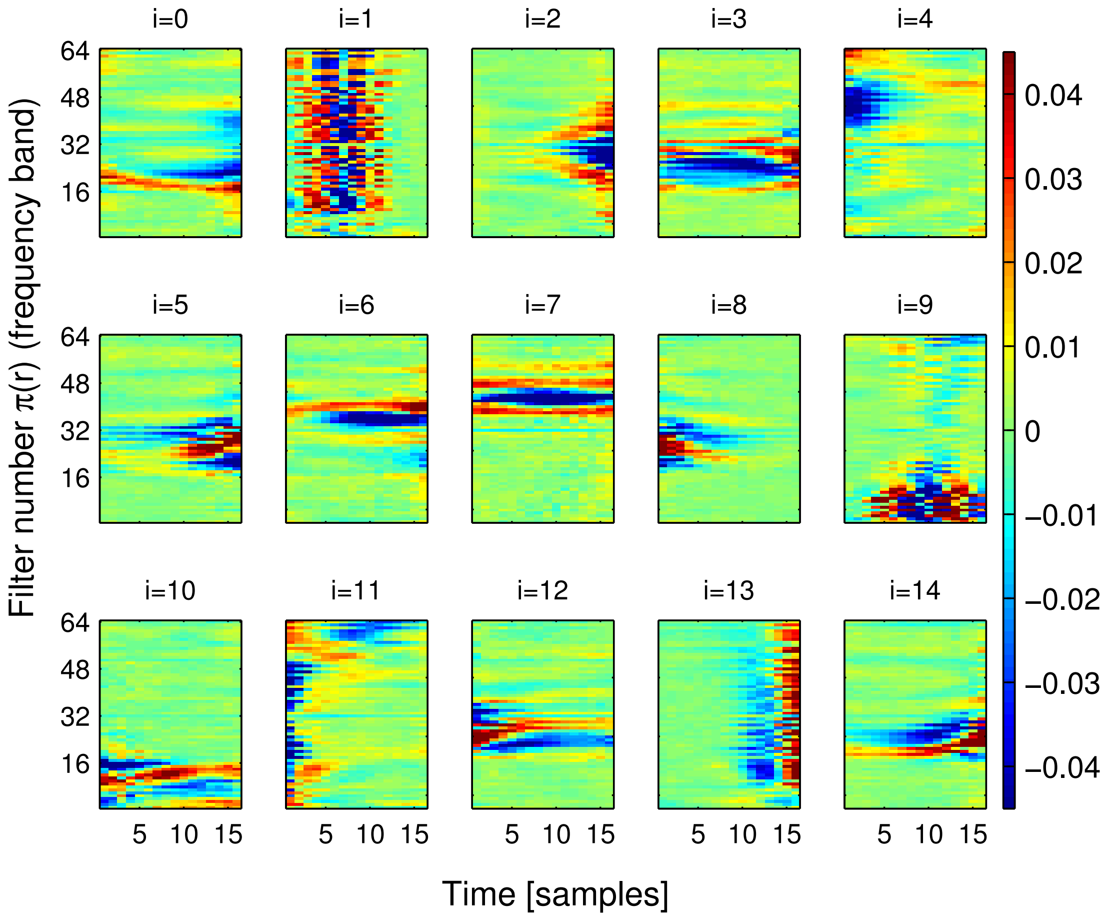
Learned Weights: First Convolutional Layer



- the (reordered) transfer functions derived from the trained convolutional filters of the first layer clearly resemble **critical bands**

Learned weights: second convolutional layer

- reordered weights of some of the 128 filters i in the 2nd convolutional layer
- vertical: frequency axis, horizontal: time axis
- dynamic patterns in both time and frequency



Outline

Experimental Results and Discussion

- training on **raw time signal** works surprisingly well
- convolutional layers improve ASR performance over fully-connected layers
- non-stationary patterns can be captured precisely
- first and second layer weights can be interpreted as filters in time/frequency

model	input	WER [%]
DNN	MFCC	22.1
	raw time signal	25.5
CNN		23.4

- the gap to MFCC's performance reduces from 15% to 6% relative WER

- for sufficient amounts of training data, models trained on the raw time signal can even outperform standard preprocessing, even for multichannel scenarios [Sainath & Weiss+ 2015]

Outline

DNN-based Single Channel Denoising for ASR

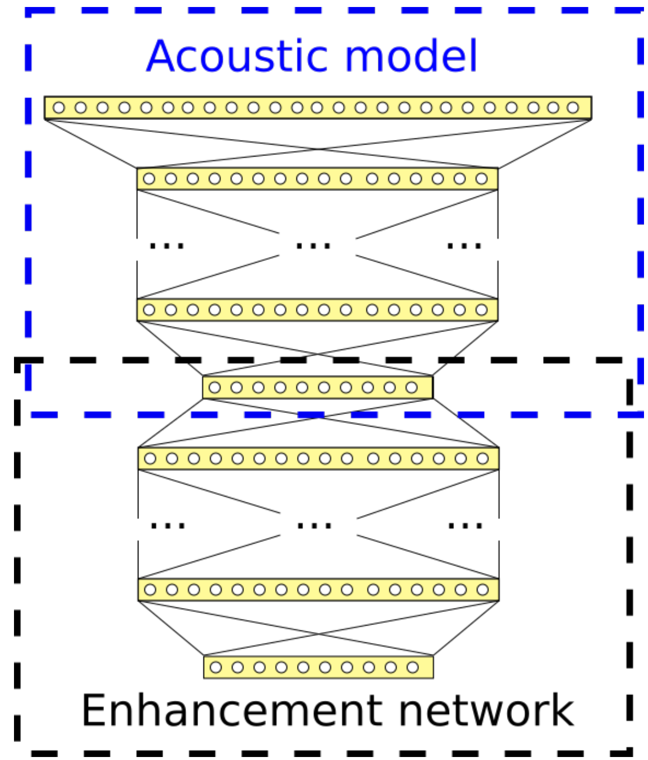
Approach:

- mapping from noisy log-mel power spectrum to clean log-mel power spectrum as e.g. done in [Xu & Du+ 2015]
- training requires two recording channels: noisy *and* clean
- e.g. MMSE loss function for DNN with linear output layer:

$$L = \frac{1}{N} \sum_{n=1}^{N} ||\hat{X}_n - X_n||_2^2$$

with

- N: the number of samples of the (mini) batch
- X_n: reference/clean log-mel power spectrum for sample n
- $\hat{X}_n$: output of enhancement network for sample n



Advantage of enhancement approach:

- can easily be combined with acoustic model for joint training, e.g. [Gao & Du+ 2015]

Outline

Model:

- multichannel speech signal with additive noise in frequency domain
- with:
  - microphone index $m = 1, \ldots, M$,
  - frame index $t = 1, \ldots, T$, and
  - frequency bin index $k = 1, \ldots, K$

$$X_m(t,k) = S_m(t,k) + N_m(t,k)$$

Filter and sum beamforming:

- ${F_m}^*(t,k)$ are the complex conjugate FIR filter coefficients applied to the m^{th} microphone:

$$Y(t,k) = \sum_{m=1}^{M} {F_m}^*(t,k) \cdot X_m(t,k)$$

Filter matrix computation:

- here for the example a GEV-beamformer [Warsitz & Haeb-Umbach 2007]
- GEV-beamformer maximizes output SNR for every frequency bin separately

$$F_m(t,k) = P\{{\mathbf{\Phi}_{\mathbf{NN}}}^{-1}(k)\mathbf{\Phi}_{\mathbf{XX}}(k)\}_m$$

  - $\mathbf{\Phi}_{\nu\nu}$ denotes the cross power spectral density matrices of signal $\nu \in \{N, X\}$
  - $P\{\cdot\}$ yields the principal component,

Utilization of noise and speech masks:

$$\mathbf{\Phi}_{\nu\nu} = \sum_{t=1}^{T} M_{\nu}(t,k)\mathbf{X}(t,k)\mathbf{X}(t,k)^{H}$$

- M_{ν}: signal masks for noise and speech
- $\mathbf{X}(t,k) = [X_1(t,k), \ldots, X_M(t,k)]^T$.

Mask estimation:

- neural networks like BLSTMs can be used for mask estimation, e.g. [Heymann & Drude+ 2015]
- this approach can be similarly applied to MVDR beamforming [Higuchi & Ito+ 2016]

Multichannel processing for ASR on raw waveform [Sainath & Weiss+ 2015]

- filters applied to time signal are learnable.
- convolutional long short-term memory deep neural network (CLDNN) jointly used for feature extraction and acoustic modeling.

Outline

Outline

Multilingual MLP Features

[Tüske & Schlüter+ 2013]

Exploitation of language independent information is viable:

- cross-lingual application of MLP features can improve performance [Stolcke & Grézl+ 2006].
- training MLP on target language usually better for similar amount of training data.

Training MLPs on multiple languages:

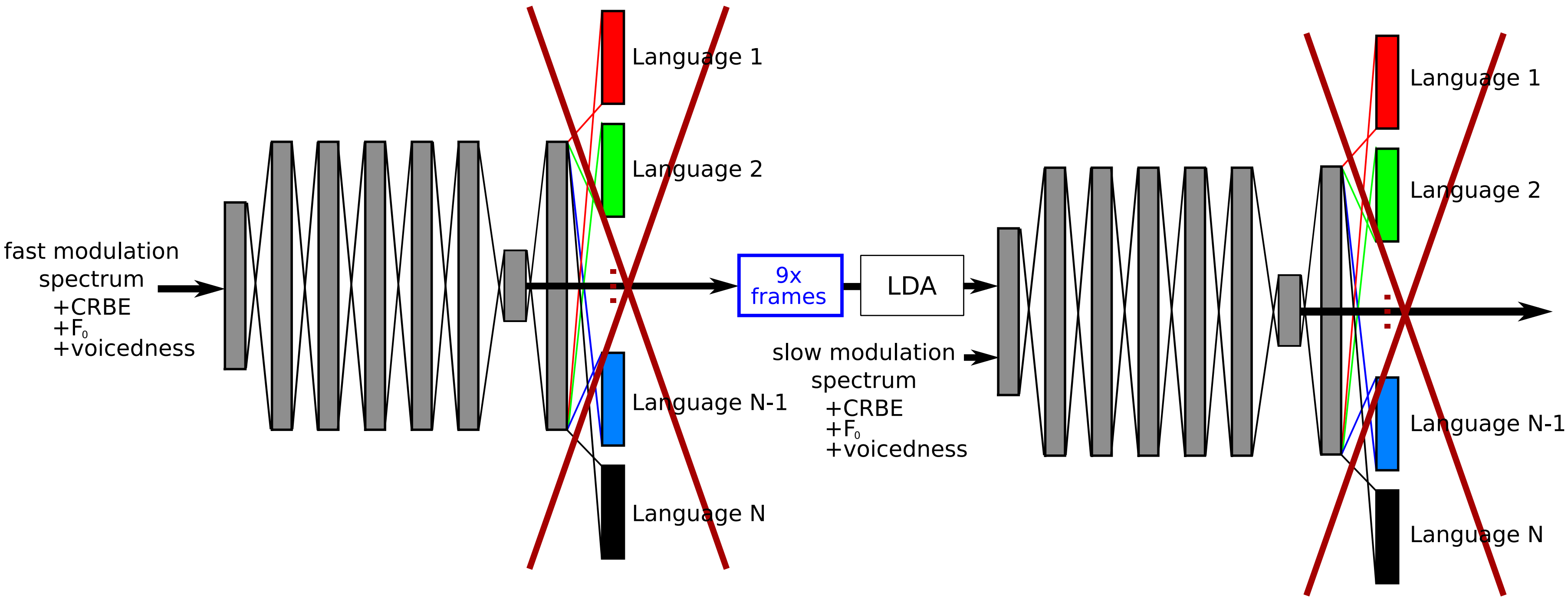
- spoken languages are based on the same speech production mechanisms.
- allows parameter sharing between languages.
- idea: share common bottleneck layer for multiple languages.
- robust feature: better portability to new language.
- exploits data available in other/multiple languages.
- serves as initialization prior to additional language specific training/fine-tuning.

Multilingual Bottleneck MLP

Handling multiple targets:

- phone set incl. **language ID** [Grézl & Karafiát+ 2011]:
  - NN also has to learn language identification.
- mapping to **common phone set** [Schultz & Waibel 2001]:
  - knowledge based (e.g IPA, SAMPA):
    often ambiguous due to simplified lexicons.
  - data-driven.
- **language dependent output layer** [Scanzio & Laface+ 2008]:
  - no need to map phonetical units to common set.
  - error back-propagation only from the active output.
  - related to multi-task training.

Architecture of Multilingual Hierarchical Bottleneck MLP



Outline

Experiments - Quaero, Small Scale

Experimental setup

- target task: French.
- 50h of speech per language (balanced corpus size)
- data available for French (FR), English (EN), German (DE), Polish (PL)
- tandem/bottleneck approach
- GMM: 4500 tied-states for each language
- shallow BN-MLPs (7000,60,7000), with phoneme targets
- speaker independent WER reported on Eval11

Effect of number of languages: the more languages, the better:

training languages				WER
FR	EN	PL	DE	[%]
✓				22.2
✓			✓	21.6
✓		✓	✓	21.5
✓	✓	✓	✓	**21.1**

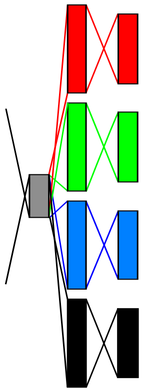
Effect of Multi- and Unilingual Bottleneck Features

input features	WER [%] for languages: FR	EN	DE	PL
MFCC	25.5	31.6	25.0	18.9
$+\mathrm{BN}_{\mathrm{uni}}$	22.2	26.8	21.3	15.7
$+\mathrm{BN}_{\mathrm{multi}}$	21.1	24.9	20.1	15.4

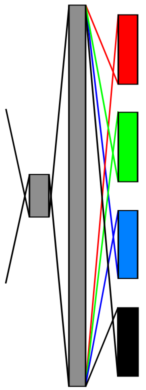
- all languages benefit from multilingual bottleneck features $\mathrm{BN}_{\mathrm{multi}}$.
- 2–5% rel. improvement over unilingual features $\mathrm{BN}_{\mathrm{multi}}$.
- 17-21% overall rel. improvement over MFCC baseline.

Experiments - Quaero, Large Scale

- speaker adaptative training.
- unbalanced corpus sizes for languages: 100h to 300h.
- deep NN structure and context-dependent NN targets.
- tuning the language dependent part of the MLP:
  - language dependent hidden layer
    - increases no. of parameters, but same training time
    - last layer: huge, but **block diagonal** weight matrix (8000x6000)



  - large, but common hidden layer
    - increases no. of parameters even further, slower training
    - last layer: huge **full** weight matrix (8000x6000)



Experiments - Quaero, Large Scale

	WER [%] for languages:			
intput features	FR	EN	DE	PL
MFCC	21.6	26.4	21.4	15.9
$+\text{BN}_{\text{uni}}$	17.3	19.7	17.2	12.3
$+\text{BN}_{\text{multi}}$	17.0	19.2	16.3	12.1
$+\text{deep BN}_{\text{uni}}$	16.7	18.8	16.8	12.1
$+\text{deep BN}_{\text{multi}}$	16.2	18.1	15.7	11.7
w/lang. dep. hidden layer	16.3	18.2	15.7	11.7
w/large lang. indep. hidden layer	16.0	17.7	15.4	11.7

- multilingual always outperform monolingual model.
- deep structure increases margin between uni- and multilingual:
  relative improvement in WER: shallow BN: 2–5%, **deep BN: 3–7%**.
- 25–30% rel. WER impr. over speaker adaptive MFCC baseline.

Multilingual Hybrid NN: Quaero English

- hybrid NN acoustic model with recent improvements.
  - 50 dim. gammatone input features, 17 frames context.
  - 12 hidden layers, 2000 nodes each.
  - activation function: rectified linear units.
  - low-rank factorized 12k output using 512 dim. linear BN.
  - WER reported on Quaero Eval corpus, 250h training data.

	Model	Criterion	WER [%]
unilingual	GMM	MPE	26.2
	hybrid NN	MPE	16.2
multilingual	hybrid NN	CE	17.3
	+fine-tuning	CE	16.7
		MPE	15.6

- initial multilingual hybrid NN results w/o further training.
- fine tuning: further optimization on target data.
- still ∼4% rel. improvement by multilingual training.

Outline

Effect of multilingual initialization

- limited amount of data available for new target language
- multilingual bottleneck (BN) MLP features:
  - 11 (non-target) languages (overall ∼800 hours of speech)
  - fine-tuned to target language.
- target language Tok Pisin, amount of training data:
  - full language pack (FLP): 40h
  - very limited language pack (VLLP): 3h

training data used for: BN features language(s)	BN features data	GMM language	GMM data	semi-supervised	LM data	TER [%]	MTWV
target	VLLP (3h)	target	VLLP (3h)	no	VLLP	56.4	0.250
	FLP (40h)					49.6	0.305
				yes		47.4	0.337
multi (11)	FLP (800h)			no		47.4	0.331
				yes		44.9	0.379
					+ web	**44.3**	**0.400**
target	FLP (40h)	target	FLP (40h)	no	FLP	40.5	0.458

Overview of OP2 results

- FLP results include:
  - Only 40h transcribed speech
- VLLP results include:
  - 3h transcribed speech
  - multilingual initialization
  - fine-tuning
  - Semi-supervised training
  - Web data

Lang. Pack	Kurmanji		Tok Pisin	
	TER [%]	MTWV	TER [%]	MTWV
FLP	65.6	0.289	40.5	0.458
VLLP	69.6	0.249	44.3	0.400

Lang. Pack	Cebuano		Kazakh	
	TER [%]	MTWV	TER [%]	MTWV
FLP	58.1	0.408	57.5	0.406
VLLP	60.3	0.354	59.9	0.411

Language Pack	Telugu		Lithuanian		Swahili	
	TER [%]	MTWV	TER [%]	MTWV	TER [%]	MTWV
FLP	70.6	0.330	50.8	0.549	44.7	0.559
VLLP	74.0	0.279	52.9	0.549	51.4	0.492

Outline

Outline

Motivation

End-to-end model:

- consistence of modeling, training, and decoding.
- cover segmentation problem by NN structure:
  sequence length, duration, and positioning of words are unknown.
- context dependence needs to be modeled.

Ultimate goals (not fully achieved yet):

- integration of NN models into Bayes decision rule.
- separation of acoustic & language model (resources usually differ).
- consistence between decision rule, evaluation measure,
  and training objective.

Review: Hidden Markov Modeling

- models words/word sequences by HMM state sequences
- within Bayes decision rule:

$$
\begin{aligned}
\arg\max_{N,w_1^N} p(w_1^N)\cdot p(x_1^N|w_1^N) &= \arg\max_{N,w_1^N} p(w_1^N)\cdot \sum_{s_1^T:w_1^N} p(x_1^T, s_1^T|w_1^N) \\
&= \arg\max_{N,w_1^N} p(w_1^N)\cdot \sum_{s_1^T:w_1^N} \prod_{t=1}^{T} p(x_t|x_1^{t-1}, s_1^t)\cdot p(s_t|x_1^{t-1}, s_1^{t-1}) \\
&= \arg\max_{N,w_1^N} p(w_1^N)\cdot \sum_{s_1^T:w_1^N} \prod_{t=1}^{T} p(x_t|s_t)\cdot p(s_t|s_{t-1}) \quad \text{1}^{\text{st}} \text{ order Markov} \\
&\approx \arg\max_{N,w_1^N} p(w_1^N)\cdot \max_{s_1^T:w_1^N} \prod_{t=1}^{T} p(x_t|s_t)\cdot p(s_t|s_{t-1}) \quad \text{Viterbi approx.}
\end{aligned}
$$

Review: Hidden Markov Modeling

Discussion:

- HMM-based standard decision rule:

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

  - in practice: **maximum** over segmentations, **especially in search** (Viterbi approximation)
  - ideally: sum over segmentations.

- inconsistency for (hybrid) NN integration into acoustic model:

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

  - NN provides state posterior, but state cond. probability needed.
  - $p(s)$ approximated, e.g. [Manohar & Povey$^+$ 2015].

Review: Hidden Markov Modeling

Discussion:

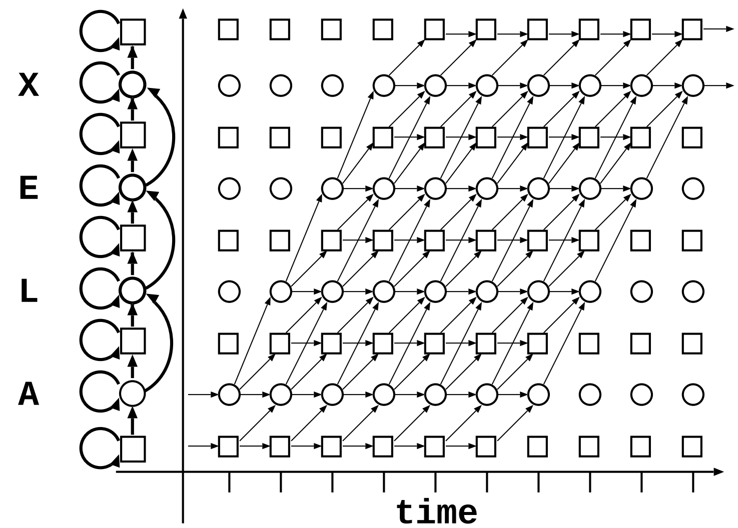
- assumption of independence of acoustic context:
  - can be relaxed by considerung window around each time frame t: $x_{t-\delta}^{t+\delta}$
  - hybrid modeling: emission probability modelled by rescaled state posteriors $p(s|x_t)$
  - observation here appears in condition only and may be replaced by full acoustic context:
    $\rightarrow$ $p(s|t, x_1^T)$ (e.g. obtained by bi-directional recurrent modeling).
- segmentation/alignment of observations to HMM states:
  - stochastic: ideally sum over all aligments.
  - explicit in case of Viterbi approximation: maximizing alignment.
- integration of language model:
  - clearly defined, can be trained separately
    (text data vs. transcribed acoustic data).
  - however, language model scaling exponent statistically unclear.
  - open issue: interaction of context dependence on observation and symbol/word level.

Outline

Alternative approach to handle segmentation problem:

- originally introduced for handwriting recognition [Graves & Fernández+ 2006, Graves & Liwicki+ 2008]
- frame-wise classification, use of LSTMs
- introduces 'blank' symbol: non-classification
- example: segmentations of the the word "ALEX":

```
AAAAALLEEEEEXXXX
A---LLL-EE----XX
-----A--L-E--X--     =    ALEX
--ALEX----------
A--------LE----X
```



- similar to 2-state HMM with globally pooled second state
- no transition model: independence assumption on symbol level
- training: from scratch, sum over all segmentations
- use of CTC in large vocabulary recognition: similar to hybrid

Contrast: What is Different from Hybrid HMM?

Where do CTC and Hybrid HMM differ?

- training criterion
- realignment in training
- alignment topology
- use of transition probabilities
- use of state priors
- NN models

CTC:	Hybrid HMM/NN:
uses Baum-Welch (full-sum)	Viterbi (maximum approximation)
realignment rate: every mini-batch	realignment rate: not at all, calculated with earlier model
topology: 1-state HMM and optional blank symbol	topology: 3-state HMM
no transition probabilities	transition probabilities
no state prior probabilities	state prior probabilities
connected to LSTM modeling	DNN/LSTM/CNN

Outline

End-to-End Approach

Motivation: End-to-end trainable neural network recognizer

- consistently integrate input and output sequences
- does not need explicit segmentation
- avoids Markov and independence assumptions

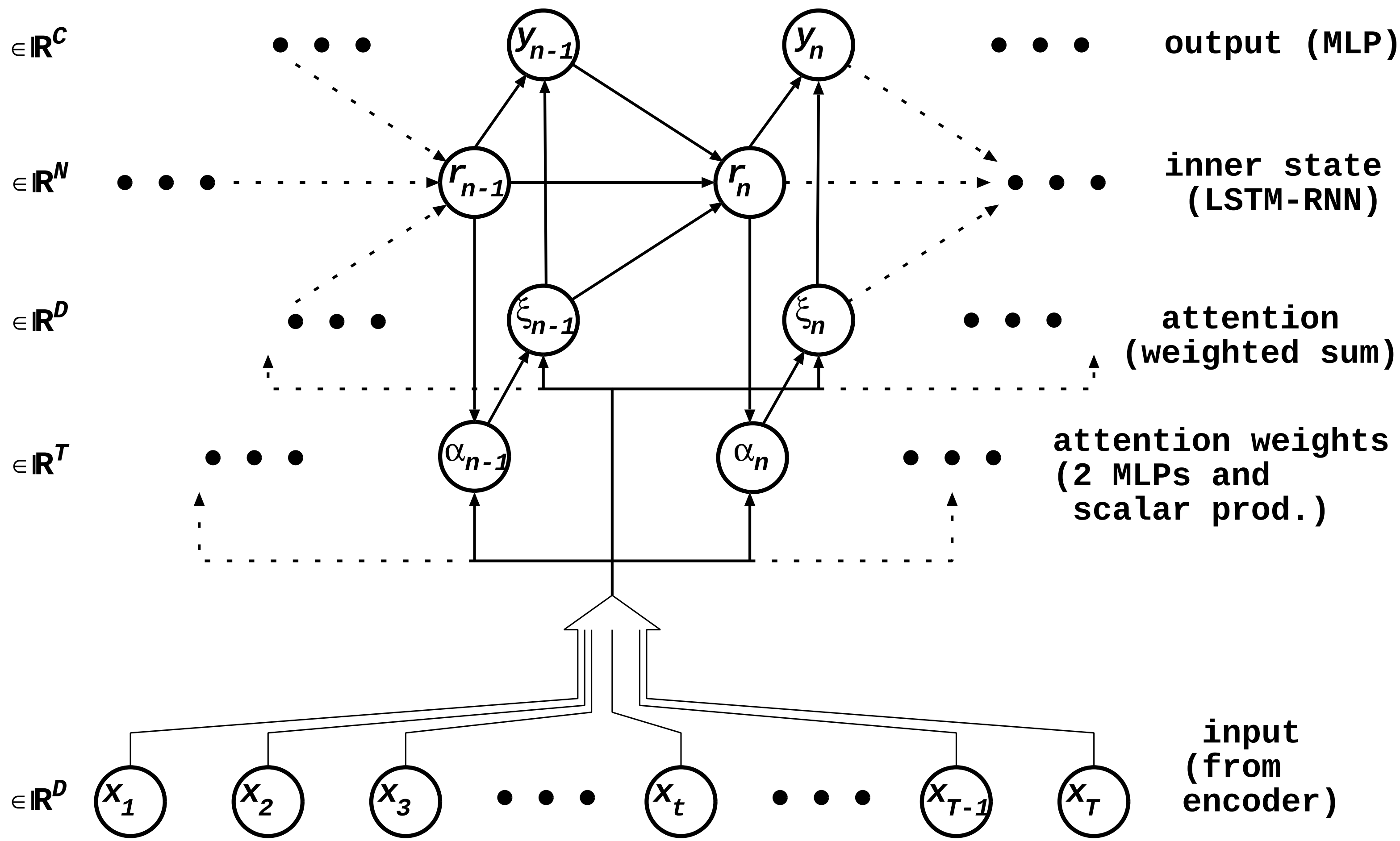
Sequence-to-sequence modeling [Sutskever & Vinyals+ 2014]:

- idea: separate processing of input and output into two models:
- **encoder**: Read the inputs and generate discriminative features
- **decoder**: Write the output symbol sequence label by label considering all encoded features

Encoder can be viewed as non-linear transformation of input:

- similar to tandem in hybrid approach (hierarchical model)
- **however**: encoder output is not related to specific output labels, as in hybrid approach
- jointly trained within the complete end-to-end structure

End-to-End Approach "Listen, Attend and Spell" [Chan & Jaitly+ 2015]



End-to-End Approach

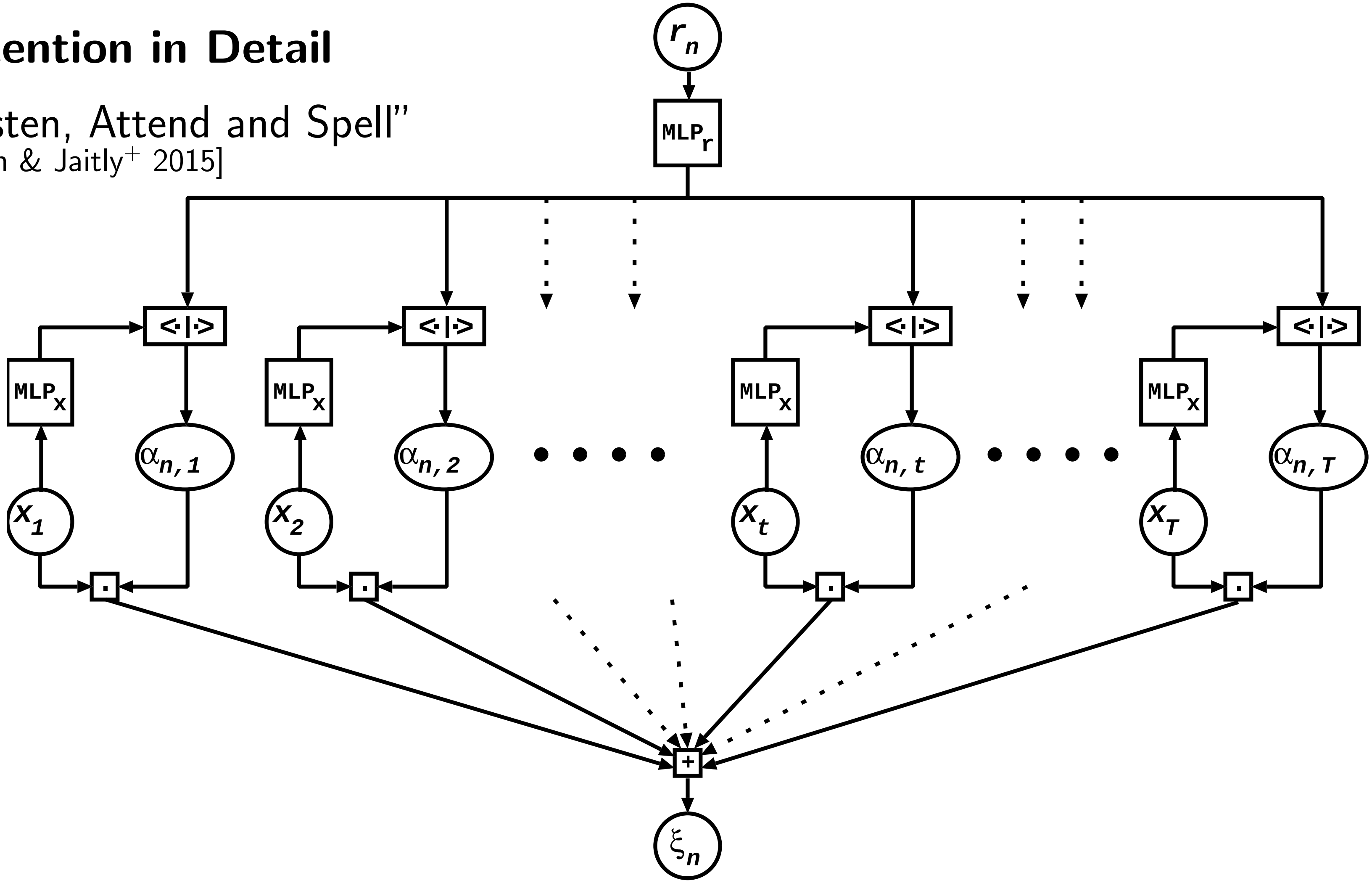
"Listen, Attend and Spell" [Chan & Jaitly$^+$ 2015]

Approach:

1. **"Listen"**:
   i. Encode input (bidirectional recurrent (LSTM) network, omitted in figure).
   Encoding usually includes gradual temporal subsampling/integration.
2. **"Attend"**: at each output symbol position n:
   i. Compute the current inner state value r_n from previous state r_{n-1}, output y_{n-1}, and expected input ξ_{n-1} from attention.
   ii. Compute attention weights $\alpha_n = attend(r_n, \ldots)$ from current state r_n and further input (see next slide).
   iii. Compute expected network input ξ_n as linear combination of input sequence x_1^T weighted by $\alpha_{n,1}^T$
3. **"Spell"**:
   i. Recurrently classify characters (symbols) from current state r_n and input ξ_n from attention.

Attention in Detail

"Listen, Attend and Spell"
[Chan & Jaitly+ 2015]



Outline

Discussion

Attention process:

- controls the segmentation
- (soft) alignment between symbol position and observations.

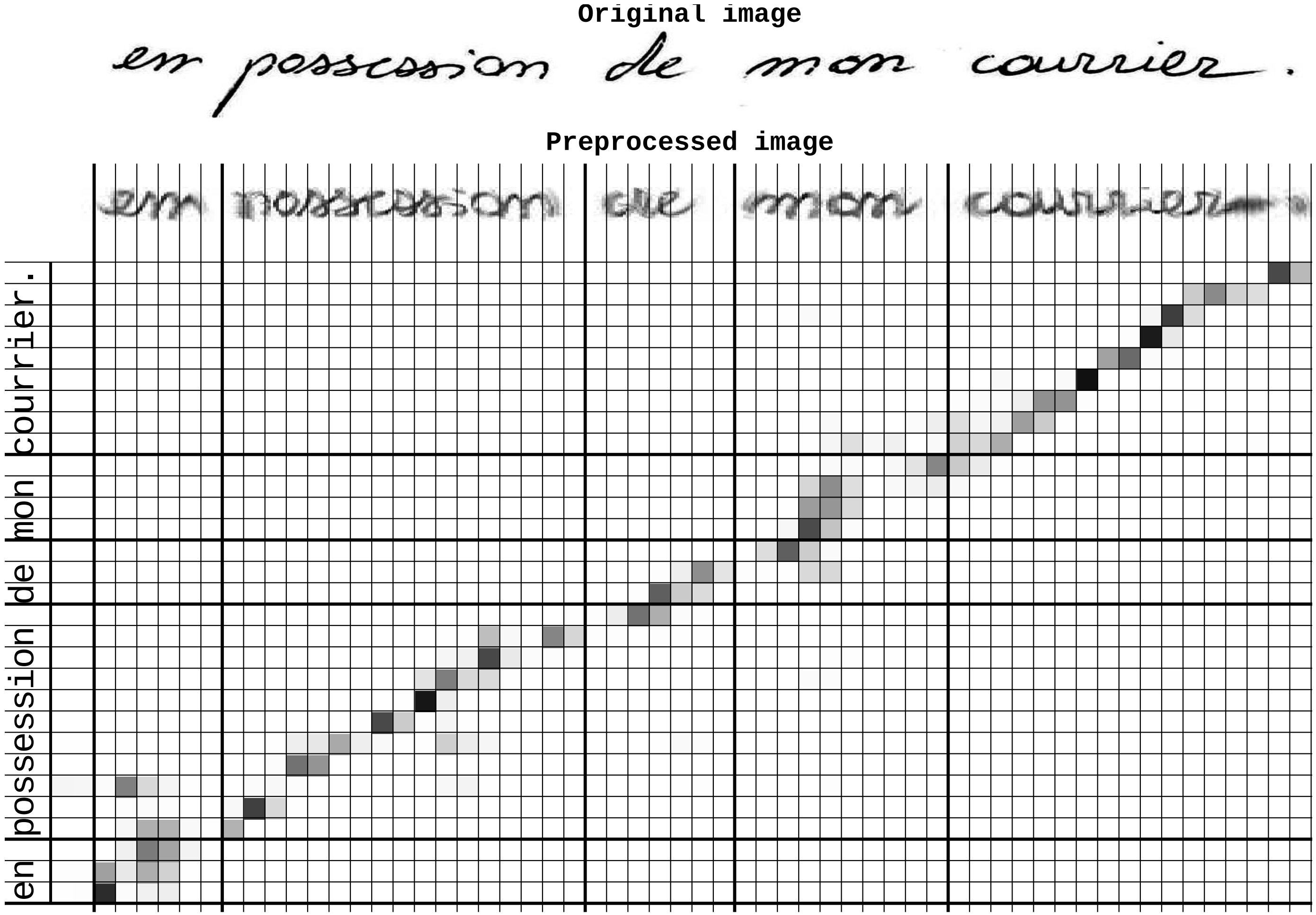
Dependencies of attention process still are an open research issue, e.g.:

- [Chan & Jaitly+ 2015] ("Listen, Attend and Spell"): $\alpha_n = attend(r_n, x_1^T)$
- [Bahdanau & Chorowski+ 2015]: $\alpha_n = attend(r_{n-1}, y_{n-1}, \xi_{n-1})$

Properties:

- no explicit alignment to specific input vectors needed.
- however, attention is **determined** by context, i.e. it is not handled as an independent hidden stochastic variable.
- as a consequence, suboptimal attention results (misalignments) cannot be rectified in the subsequent search process, as in HMM based modeling.

Attention Modeling Example from Handwriting



Sequence-to-Sequence Approach

Results: RIMES Offline Handwriting Recognition

- input: 8×32 image slices resulting from sliding window (shift 3).
- input layer: CNN with filter size 3×3 and 64 features, no pooling.
- hybrid: 4 BLSTM layers with 512 cells in each direction,
  - realignment: retraining on new alignment created based on hybrid.
- attention-based: encoder (almost) equal to hybrid:
  - "subsampling" by factor of 2 after 2[nd] and 4[th] BLSTM layer (stacking) (no subsampling/stacking in framewise system).
- decoder network: single BLSTM with 512 cells for each direction.
- # params: $\sim$ 20.8M for encoder/hybrid +700k for decoder BLSTM.

approach	WER [%]	CER [%]
hybrid HMM	13.0	7.6
+ realignment	12.9	5.8
attention-based	16.2	8.0
+ LM rescoring	14.2	6.3

Outline

Inverted HMM Derivation

- neural network based modeling provides HMM state posteriors.
- can (sub)word sequences directly be modeled using state posteriors?
- idea: **invert** alignment problem:
  - state boundaries t_1^N as hidden variables,
  - (triphone state) label sequence α_1^N directly represents word (sequence) template.
  - approach: alternative decomposition by chain rule/Bayes identity:

$$\begin{aligned} p(\alpha_1^N|x_1^T) &= \sum_{t_1^N} p(\alpha_1^N, t_1^N|x_1^T) \\ &= \sum_{t_1^N} p(\alpha_1^N|t_1^N, x_1^T) \cdot p(t_1^N|x_1^T) \\ &= \sum_{t_1^N} \prod_{n=1}^{N} p(\alpha_n|\alpha_1^{n-1}, t_1^N, x_1^T) \cdot p(t_n|t_1^{n-1}, x_1^T) \\ &= \sum_{t_1^N} \prod_{n=1}^{N} \underbrace{p(\alpha_n|\alpha_1^{n-1}, t_{n-1}, t_n, x_1^T)}_{\text{NN-based posterior}} \cdot \underbrace{p(t_n|t_{n-1})}_{\text{length model}} \end{aligned}$$

Inverted Search

Discussion:

- **inverted search**, as times are aligned to triphone (state) labels, instead of vice versa.

$$p(\alpha_1^N|x_1^T) = \sum_{t_1^N} \prod_{n=1}^{N} \underbrace{p(\alpha_n|\alpha_1^{n-1}, t_{n-1}, t_n, x_1^T)}_{\text{NN-based posterior}} \cdot \underbrace{p(t_n|t_{n-1})}_{\text{length model}}$$

- symbol by symbol hypothesis generation.
- language model integrated into state posterior.

Proof of concept:

- RIMES isolated word handwritten character recognition task [Doetsch & Hegselmann+ 2016]

model	WER [%]	CER [%]
hybrid HMM	7.1	3.0
CTC	6.7	2.8
attention	7.7	4.1
inverted HMM	7.5	2.9

Inverted Search: Experiments

First speech recognition results (ongoing work):

- CHiME-4 speech separation and recognition challenge [Doetsch & Hannemann+ 2017]

	WER [%]	
model	dev	eval
hybrid HMM	6.1	8.1
inverted HMM	5.7	8.8

Current research questions:

- how to model state posterior? - not necessearily the same, as in hybrid approach: here state posterior covers multiple time frames in one step.
- what length model should be used? - existing HMM based work less successful.
- where are the words? - word sequence determines state sequence: effectively states represent subwords (or even words itself!).
- how to fit in (separately trained) **language model**?

Outline

Statistical approach

- four key ingredients:
  - choice of performance measure: errors at string, word, phoneme, frame level
  - probabilistic models at these levels and the interaction between these levels
  - training criterion along with an optimization algorithm
  - Bayes decision rule along with an efficient implementation

- about recent work on artificial neural nets in the last ten years:
  - significant improvements by deep MLPs and LSTM-RNNs
  - they provide one more type of probabilistic models within classical *Bayes* framework

- properties of neural networks in the context of statistical ASR:
  - Do the NNs discover dependencies that we cannot model explicitly?
  - Is it a better way of smoothing that makes the NN better?
  - Is it the use of crossvalidation that makes NNs successful?

- long-term research topics at RWTH:
  - relation of training criteria and error rate (frame, phoneme, word, sentence)
  - open lexicon ASR: any letter sequence can be recognized
  - (fully) unsupervised training: without *any* transcribed training data

Future Challenges

- specific future challenges for statistical approach (incl. NNs) in general:
  - complex mathematical model that is difficult to analyze
  - questions: can we find suitable mathematical approximations with more explicit descriptions of the dependencies and level interactions and of the performance criterion (error rate)?

- specific challenges for artificial neural networks:
  - methods with better convergence?
  - can the HMM-based alignment mechanism be replaced?
  - can we find NNs with more explicit probabilistic structures?

- potential challenges from comparison to biological structures:
  - what connectivity is needed for speech modeling? can efferent connections contribute?
  - how to analyze large/complex networks?
  - how can neural networks lead to effictive search organization?
  - how is sequential context encoded in the human brain?
  - do we need spiking networks in ASR?
  - what neural mechanisms are required, and how to implement them efficiently in deep ANNs?
  - ...

Thank you for your attention

Any questions?

Acknowledgements

Part of the word presented in this presentation was supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Defense U.S. Army Research Laboratory (DoD/ARL) contract no. W911NF-12-C-0012. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoD/ARL, or the U.S. Government.

Corpus IDs Babel

Table: Language packs released by IARPA to the project participants. Last row in each of the four period blocks corresponds to the surprise (i.e. evaluation) language.

	Language	ID	Language pack version
BP	Cantonese	101	IARPA-babel101-v0.4c
BP	Pashto	104	IARPA-babel104b-v0.4aY
BP	Turkish	105	IARPA-babel105b-v0.4
BP	Tagalog	106	IARPA-babel106-v0.2f
BP	Vietnamese	107	IARPA-babel107b-v0.7
OP2	Kurmanji Kurdish	205	IARPA-babel205b-v1.0a
OP2	Tok Pisin	207	IARPA-babel207b-v1.0a
OP2	Cebuano	301	IARPA-babel301b-v2.0b
OP2	Kazakh	302	IARPA-babel302b-v1.0a
OP2	Telugu	303	IARPA-babel303b-v1.0a
OP2	Lithuanian	304	IARPA-babel304b-v1.0b
OP2	Swahili	202	IARPA-babel202b-v1.0d

	Language	ID	Language pack version
OP1	Assamese	102	IARPA-babel102b-v0.5a
OP1	Bengali	103	IARPA-babel103b-v0.4b
OP1	Haitian Creole	201	IARPA-babel201b-v0.2b
OP1	Lao	203	IARPA-babel203b-v3.1a
OP1	Tamil	204	IARPA-babel204b-v1.1b
OP1	Zulu	206	IARPA-babel206b-v0.1e
OP3	Pashto	104	IARPA-babel104b-v0.4bY
OP3	Guarani	305	IARPA-babel305b-v1.0c
OP3	Igbo	306	IARPA-babel306b-v2.0c
OP3	Amharic	307	IARPA-babel307b-v1.0b
OP3	Mongolian	401	IARPA-babel401b-v2.0b
OP3	Javanese	402	IARPA-babel402b-v1.0b
OP3	Dholuo	403	IARPA-babel403b-v1.0b
OP3	Georgian	404	IARPA-babel404b-v1.0a

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?"

References

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.*

References

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,*

References

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.