Anniversary Symposium *InterACT – 25 years*
Karlsruhe Institute of Technology (KIT), Baden-Baden, Germany,
July 14/15, 2016

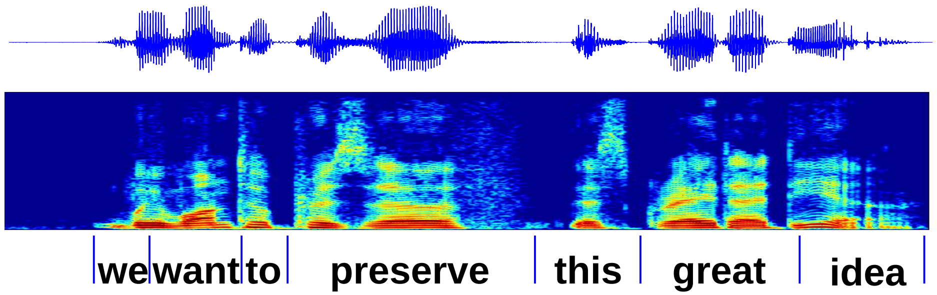
On Architectural Issues of Neural Networks in Speech Recognition

Hermann Ney
Human Language Technology and Pattern Recognition
RWTH Aachen University, Aachen, Germany

IEEE Distinguished Lecturer 2016/17

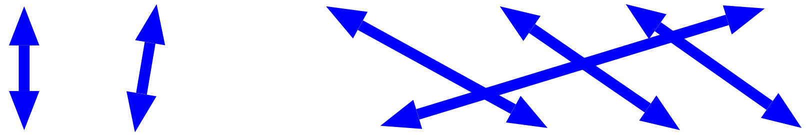
Human Language Technology (HLT)

Automatic Speech Recognition (ASR)



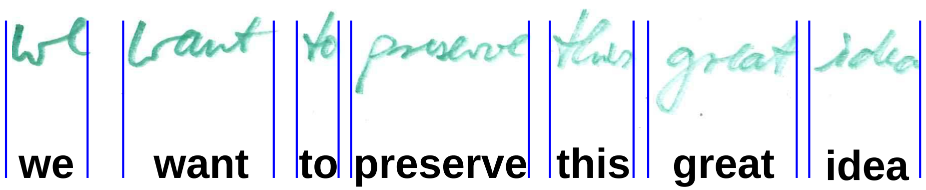
Statistical Machine Translation (SMT)

wir wollen diese große Idee erhalten



we want to preserve this great idea

Handwriting Recognition (Text Image Recognition)



tasks:
– speech recognition
– machine translation
– handwriting recognition

unifying view:
– input string $\rightarrow$ output string
– output string: natural language

RWTH's Joint Projects with InterACT: KIT, CMU or HKUST

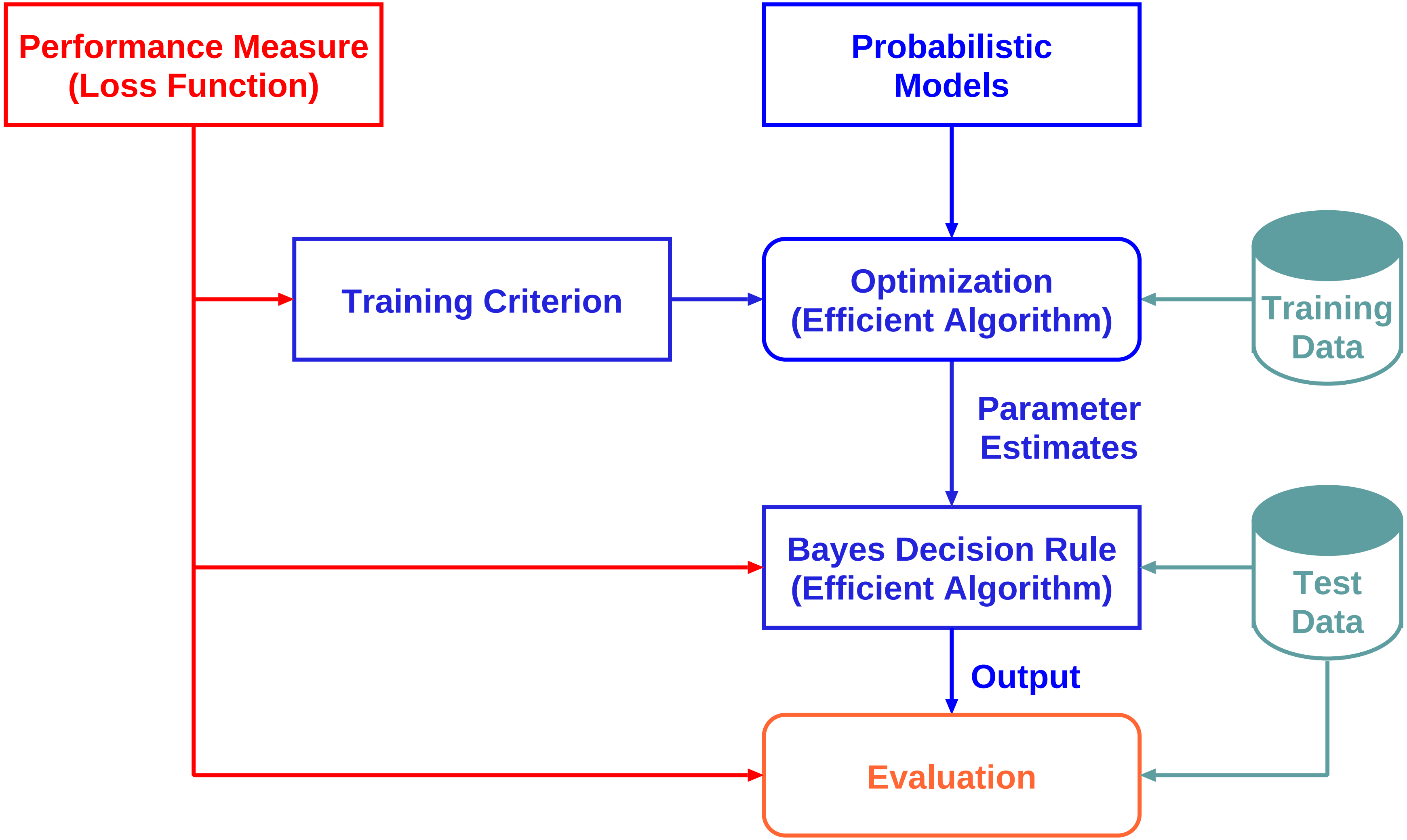
- **VERBMOBIL 1993-2000: funded by German BMBF**
  toy task (8000-word vocabulary): recognition and translation for appointment scheduling
- **TC-STAR 2004-2007: funded by EU**
  - **real-life task, open domain, large vocabulary:**
    first research system for speech translation (EU parliament)
  - **partners: KIT Karlsruhe, FBK Trento, LIMSI Paris, UPC Barcelona, IBM-US Research, ...**
- **GALE 2005-2011: funded by US DARPA**
  - **emphasis on Chinese and Arabic speech and text**
  - **largest project ever on speech and language: 40 Mio USD per year**
- **BOLT 2011-2015: funded by US DARPA**
  emphasis on colloquial text for Arabic and Chinese
- **QUAERO 2008-2013: funded by OSEO France**
  European languages, more colloquial speech, handwriting
- **EU-BRIDGE 2012-2014: funded by EU**
  emphasis on recognition and translation of lectures (TED, ...)
- **BABEL 2012-2016: funded by US IARPA**
  speech recognition for low-resource languages (and noisy audio!)

Evaluation Campaigns: InterACT (KIT)

evaluations of ASR and SMT systems:

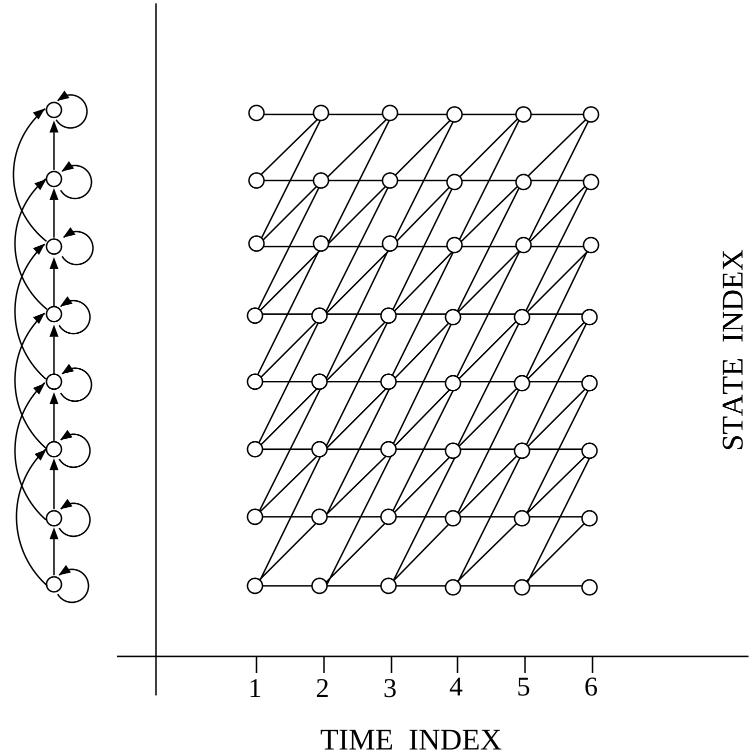
- **project related evaluations:**
  - **VERBMOBIL**
  - **TC-STAR**
  - **QUAERO**
  - **EU-BRIDGE**
- **public evaluation campaigns:**
  - **NIST/LDC/DARPA**
  - **IWSLT (organized by InterACT members)**
  - **ACL WMT**
- **joint submissions with KIT/InterACT:**
  system combination

Statistical Approach: No Alternative
(incl. Artificial Neural Networks!)



Hidden Markov Models (HMM)

- **fundamental problem in ASR: non-linear time alignment**
- **Hidden Markov Model:**
  - **linear chain of states $s = 1, ..., S$**
  - **transitions: forward, loop and skip**
- **trellis:**
  - **unfold HMM over time $t = 1, ..., T$**
  - **path: state sequence $s_1^T = s_1 ... s_t ... s_T$**
  - **observations: $x_1^T = x_1 ... x_t ... x_T$**



general view:

– two sequences without synchronization: acoustic vectors and states (with labels)

– HMM: mechanism that takes care of the synchronizatiojn (=alignment) problem

Hidden Markov Models (HMM)

The acoustic model $p(X|W)$ provides the link between word sequence 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): $(s, W) \rightarrow a = a_{sW}$**

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

  typical approach: phoneme models in triphone context:
  decision trees (CART) for finding equivalence classes

- **refinements:**
  – augmented feature vector: context window around position t
  – subsequent LDA (linear discriminant analysis)

Hybrid Approach: HMM and ANN

consider modelling the acoustic vector x_t in an HMM:

- **re-write the emission probability for annotation label a and acoustic vector x_t (strictly speaking: an approximation only):**

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

  – **prior probability $p(a)$: estimated as relative frequencies**
  – **for recognition purposes: the term $p(x_t)$ can be dropped**

- **result: model the label posterior probability by an ANN:**

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

  rather than the state emission distribution $p(x_t|a)$

- **justification:**
  – **easier learning problem: labels $a = 1, ..., 5000$ vs. vectors $x_t \in \mathbb{R}^{D=40}$**
  – **well-known result in pattern recognition/machine learning; but ignored in ASR due to the mathematical beauty of the EM algorithm**

History: ANN in Acoustic Modelling

- **1988 [Waibel & Hanazawa^{+} 88]:**
  phoneme recognition using time-delay neural networks
- **1989 [Bridle 89]:**
  softmax operation for probability normalization in output layer
- **1990 [Bourlard & Wellekens 90]:**
  – **for squared error criterion, ANN outputs can be interpreted as class posterior probabilities (rediscovered: Patterson & Womack 1966)**
  – **they advocated the *hybrid approach*: use the ANN outputs to replace the emission probabilities in HMMs**
- **1993 [Haffner 93]:**
  sum over label-sequence posterior probabilities in hybrid HMMs
- **1994 [Robinson 94]: recurrent neural network**
  – **competitive results on WSJ task**
  – **his work remained a singularity in ASR**

experimental situation:
– until 2011: ANNs were never really competitive with Gaussian mixture models
– after 2011: yes, *deep learning* [Deng & Hinton 2012]

more ANN approaches:

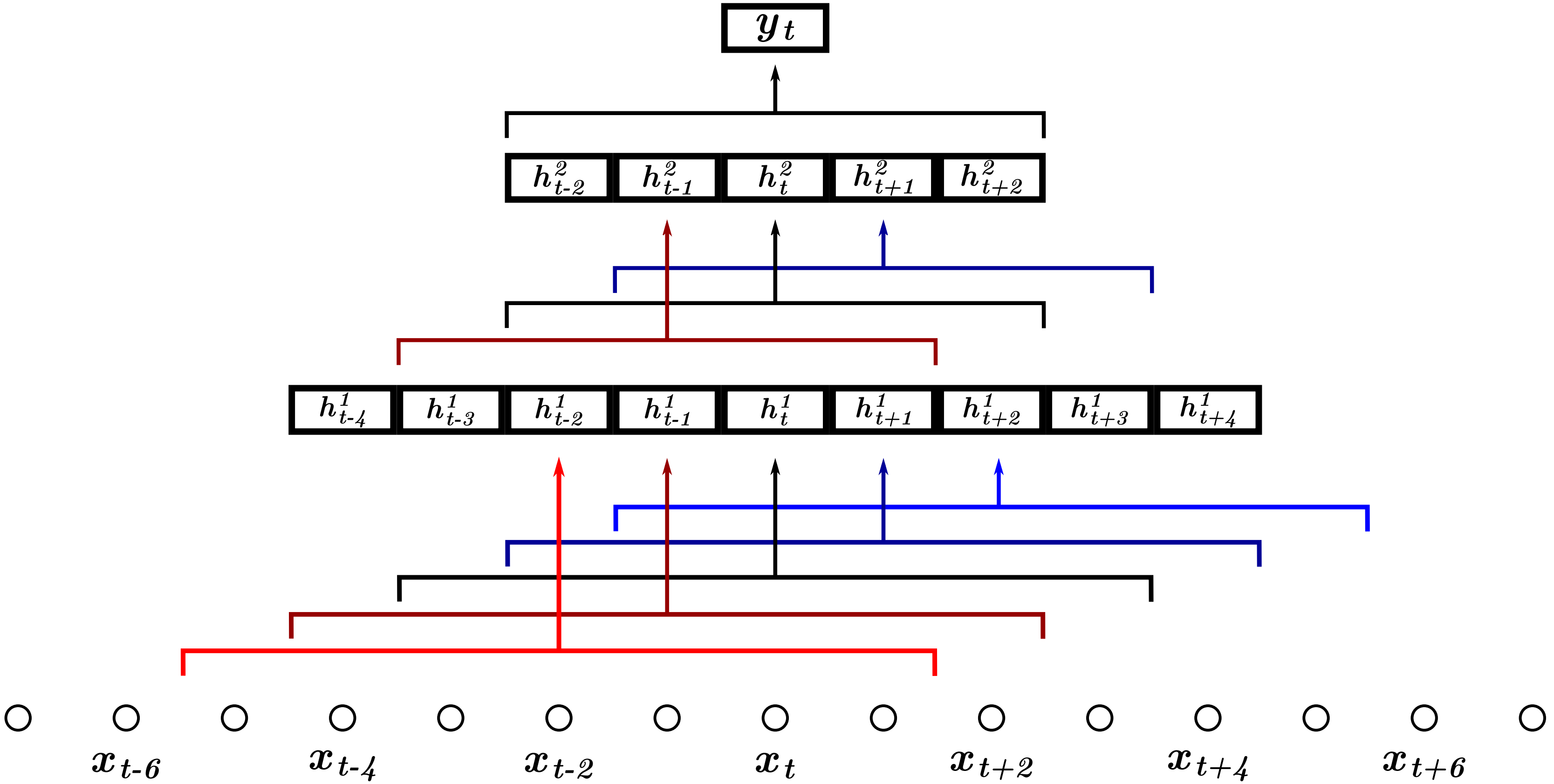
- **1994 [LeCun & Bengio$^+$ 94]: convolutional neural networks**
- **1997 A. Waibel's team [Fritsch & Finke$^+$ 97]: hierarchical mixtures of experts**
- **1997 [Hochreiter & Schmidhuber 97]: long short-term memory neural computation with extensions [Gers & Schraudolph$^+$ 02]**

renaissance of ANN: concepts of deep learning and related ideas:

- **2000 [Hermansky & Ellis$^+$ 00]: tandem approach: multiple layers of processing by combining Gaussian model and ANN for ASR**
- **2002 [Utgoff & Stracuzzi 02]: many-layered learning for symbolic processing**
- **2006 [Hinton & Osindero$^+$ 06]: introduced what he called *deep learning (belief nets)***
- **2008 [Graves 08]: good results on LSTM RNN for handwriting task**
- **2012 Microsoft Research [Dahl & Yu$^+$ 12]:**
  – combined Hinton's deep learning with hybrid approach
  – significant improvement by deep MLP on a large-scale task
- **since 2012: other teams confirmed significant reductions of WER**

TDNN: Time Delay Neural Network
[Waibel & Hanazawa+ 88]

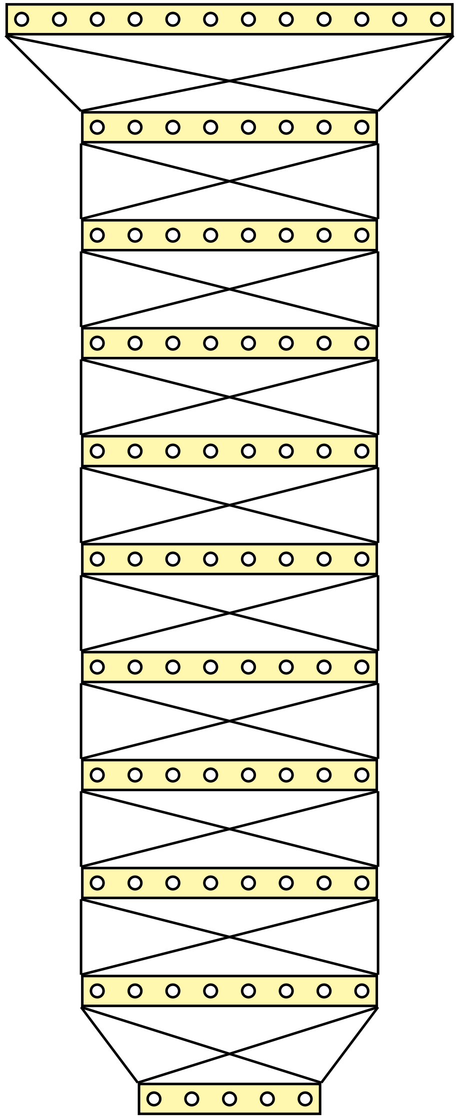
TDNN: feed-forward multi-layer perceptron with special properties:
– long temporal context
– weight sharing



TDNN: Time Delay Neural Network

- **first (?) publication: [Waibel & Hanazawa$^+$ 88] at ICASSP 1988, New York**
- **full journal paper: [Waibel & Hanazawa$^+$ 89] in**
  IEEE Transactions on Acouctics, Speech and Signal Processing 1989
  - **2036 citations (*Google Scholar*)**
  - **1116 citations on 3 more papers on TDNN 1989/90**
- **recent work by Dan Povey's team [Peddinti & Povey$^+$ 15] at Interspeech 2015: improvements over widely used deep MLP approach**
  - **on many of the standard ASR tasks (WSJ, Switchboard, Librispeech, ...)**
  - **on ASPIRE challenge (IARPA, March 2015): reverberant speech in farfield speech recognition**

Today vs. 1988-94: What is Different?



most popular and widely used:

feed-forward multi-layer perceptron (FF MLP)

– operations: matrix · vector

– nonlinear activation function

comparison for ASR: today vs. 1988-1994:

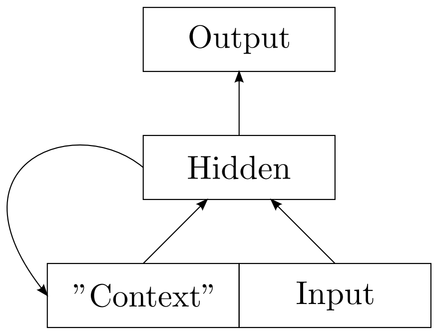
- **number of hidden layers:**
  10 (or more) rather than 2-3
- **number of output nodes (phonetic labels):**
  5000 rather than 50
- **optimization strategy:**
  practical experience and heuristics,
  e.g. layer-by-layer pretraining
- **much more computing power**

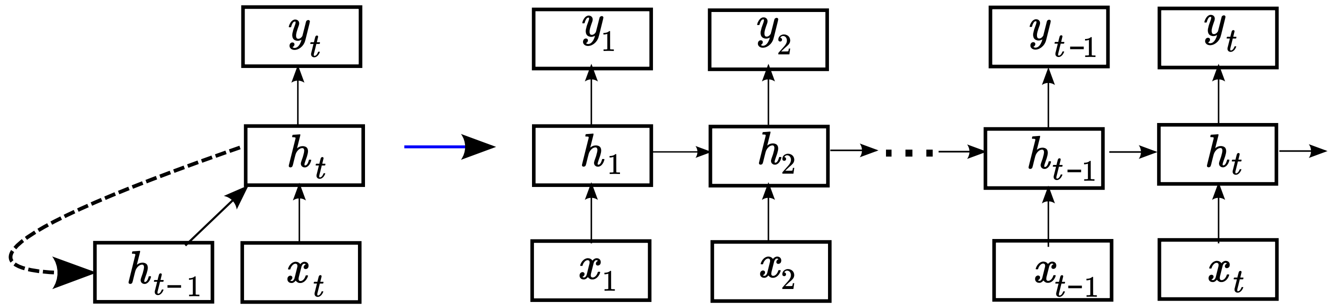
overall result:

– huge improvement by ANN

– WER is (nearly) halved !!

principle for string processing over time $t = 1, ..., T$:
– introduce a memory (or context) component to keep track of history
– quantities: input = observation x_t, memory h_{t-1}, output distribution y_t





extensions:
– bidirectional variant [Schuster & Paliwal 1997]
– feedback of output labels
– long short-term memory [Hochreiter & Schmidhuber 97; Gers & Schraudolph$^+$ 02]
– deep structure: several hidden layers

Direct Model of Label Sequence
(spirit of CTC: connectionist temporal classification)

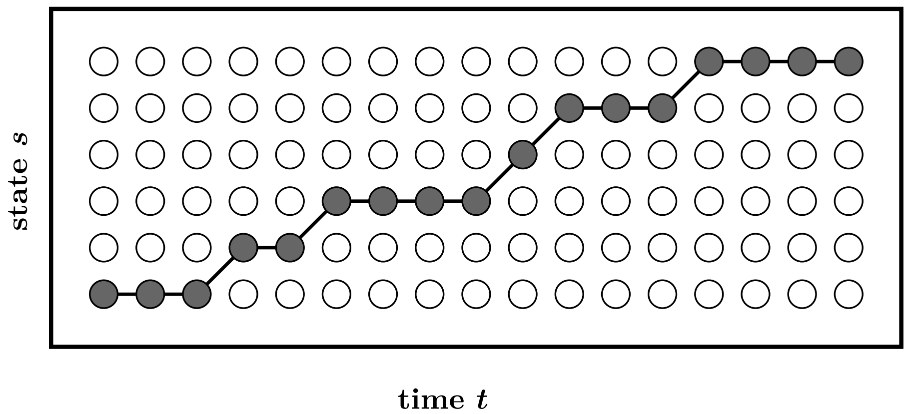
re-formulate the problem of speech recognition:

- **sequence of phonetic labels (e.g. CART):** $a_s, s = 1, ..., S$
  (which fully determines the sequence of words)
- **key quantity: (local) label posterior probability calculated by an ANN**

$$p_t(a|x_1^T) = p_t(a|x_{t-\delta}^{t+\delta})$$

- **model localization effect by alignments, i.e. mappings from time to states:**

$$t \rightarrow s = s_t$$



Direct Model of Label Sequence

sum over all hidden alignments s_1^T:

$$
\begin{aligned}
p(a_1^S|x_1^T) \;&=\; \sum_{s_1^T} p(a_1^S, s_1^T|x_1^T) \;=\; \ldots \\
&=\; \sum_{s_1^T} \prod_t p_t\big(a_{s_t}|x_1^T\big) \;=\; \sum_{s_1^T} \prod_t p_t\big(a_{s_t}|x_{t-\delta}^{t+\delta}\big)
\end{aligned}
$$

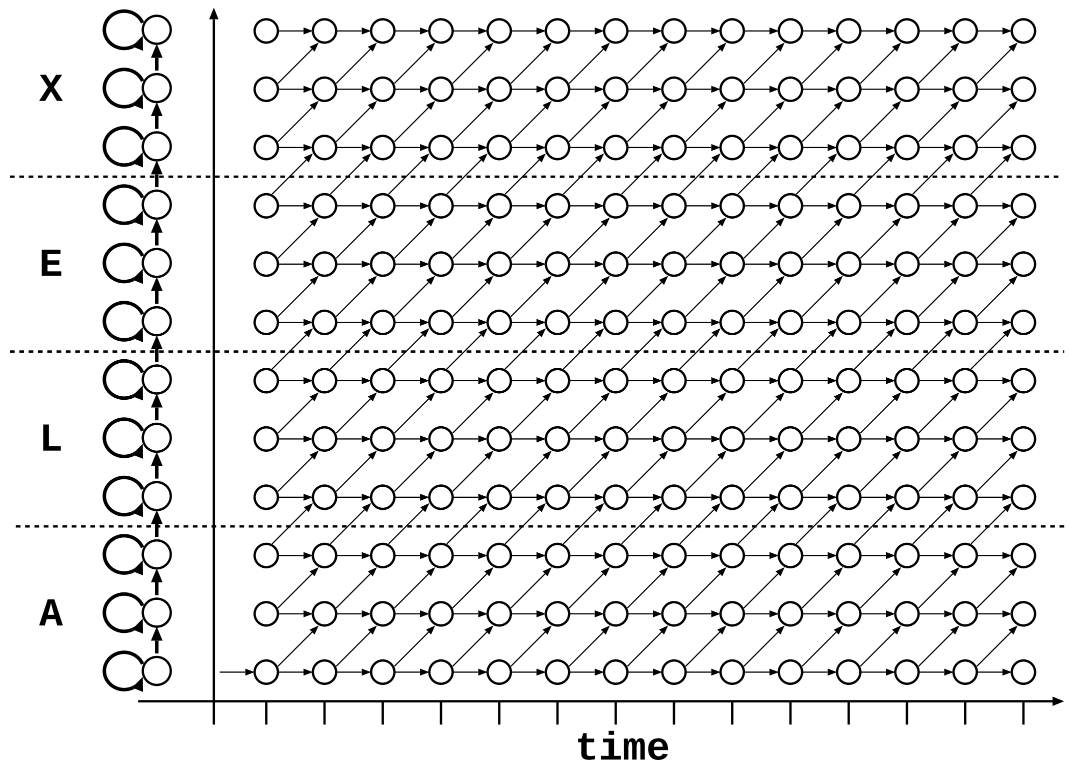
open issues:
– how to include the transition probabilities
– how to include the language model
– how to perform end-to-end training

requirement:

avoid the global re-normalization as in discriminative/hybrid HMM

Comparison with Discriminative/Hybrid HMM

- **topology:**
  conventional HMM structure
- **important differences:**
  – **no joint model $p(a_1^S, x_1^T)$**
  – **no global re-normalization (e. g. lattice)**
- **open issues:**
  – **transition probabilities**
  – **language model**
  – **consistent training criterion:**
  sum over all alignments,
  end-to-end training,...

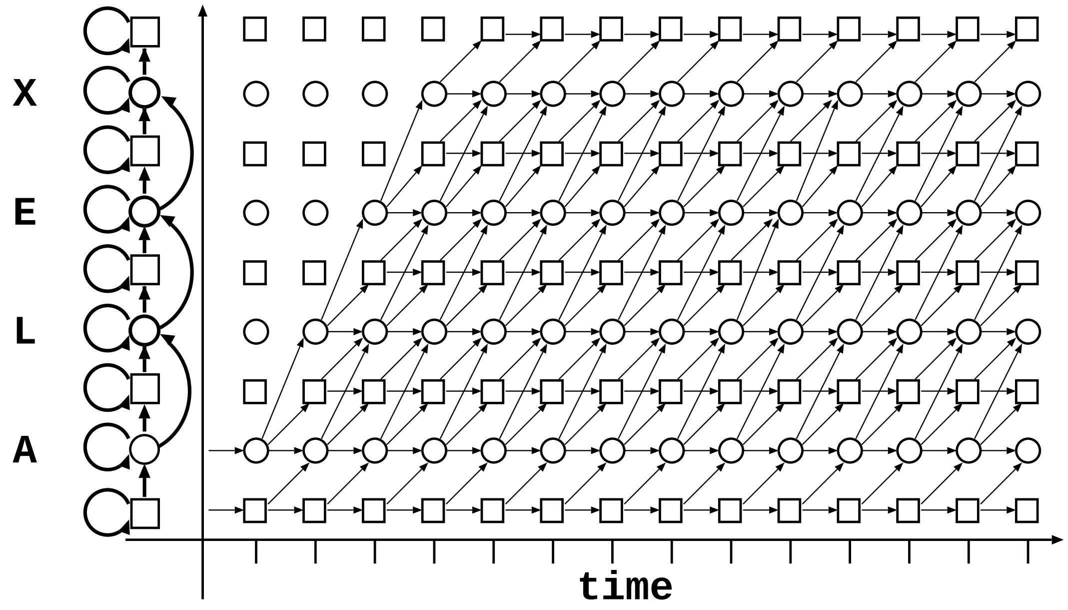


goal: avoid joint probability $p(a_1^S, x_1^T)$ as in discriminative/hybrid HMM

Comparison with CTC: connectionist temporal classification [Graves & Fernandez^{+} 06]

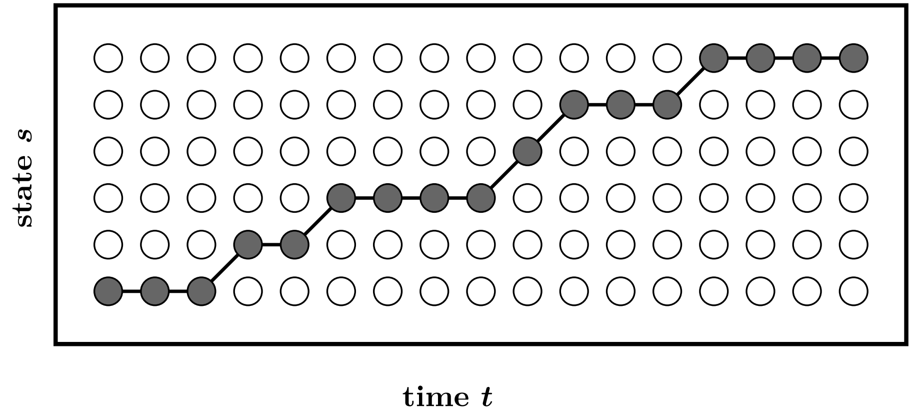
characteristic properties of CTC:

- topology: for each symbol label: single state + blank state
- no transition probabilities
- training criterion: sum
- ANN structure: LSTM RNN or ...?



experiments for CTC and related neural network approaches:
– good results reported
– reason: LSTM RNN?
– direct comparison: to be done

Direct Model of Label Sequence: Inverted Alignments



– re-interpretation of ASR: segmentation and classification problem
– consider inverted alignments, i.e. from state s to time t:

$$s \rightarrow t = t_s$$

– sum over inverted alignments as hidden variables t_1^S:

$$p(a_1^S|x_1^T) = \sum_{t_1^S} p(a_1^S, t_1^S|x_1^T) = \ldots =$$

$$= \sum_{t_1^S} \prod_{s=1}^{S} p_{t_s}(a_s|x_1^T) = \sum_{t_1^S} \prod_{s=1}^{S} p_{t_s}(a_s|x_{t_s-\delta}^{t_s+\delta})$$

experiments: underway

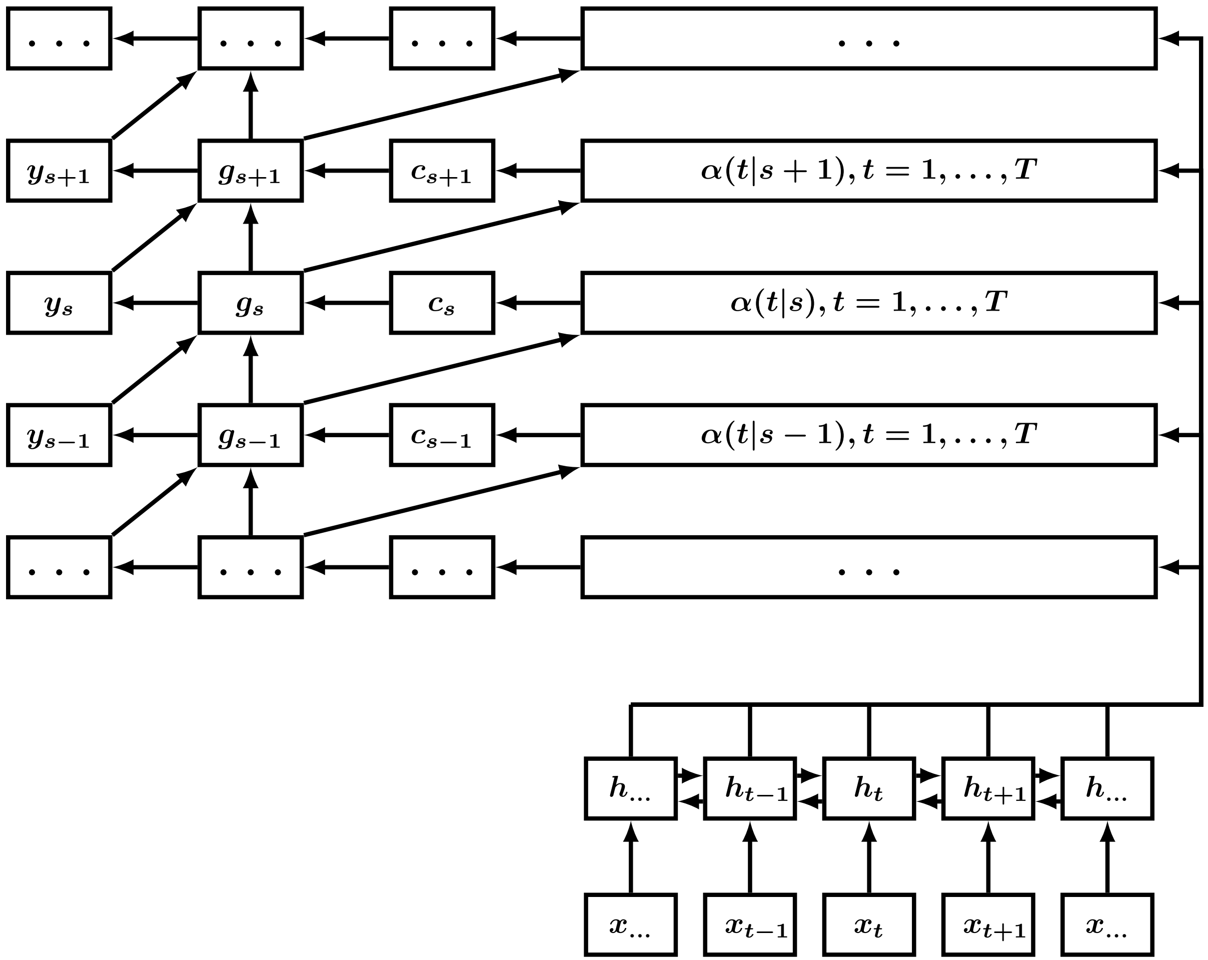
Mechanism of Attention: Alignment by ANN (originally introduced for MT [Bahdanau & Cho+ 15])

mechanism of attention:
ANN only

alignment direction:
from state s to time t

occupation probabilities:
$\alpha(t|s)$

experiments:
ongoing work, many teams



Summary

Architectural Issues of ANN in ASR Systems:

- **starting point: direct model of label sequences:**
  - **use ANN output as label posterior probability**
  - **(try to) avoid global re-normalization (no denominator/lattice)**
- **open questions:**
  - **how to include transition probabilities?**
  - **how to include language model?**
  - **end-to-end training: suitable training criterion**
- **some localization is needed: alignments**
  - **inverted alignments vs. traditional alignments**
  - **attention-based mechanism: alternative?**
- **experimental results: room for improvements**
  - **a large number of ongoing studies**
  - **clear conclusions: difficult**

Congratulations to InterACT and Alex on 25 successful years!

Best wishes for the coming 25 years!

REFERENCES

References

[Bahdanau & Cho+ 15] D. Bahdanau, K. Cho, Y. Bengio: Neural machine translation by jointly learning to align and translate. Int. Conf. on Learning and Representation (ICLR), San Diego, CA, May 2015.

[Bahl & Jelinek+ 83] 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.

[Bahl & Brown+ 86] 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.

[Beck & Schlüter+ 15] E. Beck, R. Schlüter, H. Ney: Error Bounds for Context Reduction and Feature Omission, Interspeech, Dresden, Germany, Sep. 2015.

[Bengio & Ducharme+ 00] Y. Bengio, R. Ducharme, P. Vincent: A neural probabilistic language model. Advances in Neural Information Processing Systems (NIPS), pp. 933-938, Denver, CO, USA, Nov. 2000.

[Botros & Irie+ 15] 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.

[Bourlard & Wellekens 90] 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.

[Bridle 89] 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.

[Brown & Della Pietra+ 93] 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.

[Castano & Vidal+ 93] 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.

[Castano & Casacuberta 97] 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.

[Castano & Casacuberta+ 97] 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, USA, July 1997.

[Dahl & Yu+ 12] 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.

[Devlin & Zbib+ 14] 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.

[Fritsch & Finke+ 97] 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.

[Gers & Schmidhuber+ 00] F. A. Gers, J. Schmidhuber, F. Cummin: Learning to forget: Continual prediction with LSTM. Neural computation, Vol 12, No. 10, pp. 2451-2471, 2000.

[Gers & Schraudolph+ 02] 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.

[Graves & Fernandez+ 06] A. Graves, S. FernÃ¡ndez, F Gomez, J. Schmidhuber: Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks. Int.Conf. on Machine Learning, Pittsburgh, USA, pp. 369-376, 2006.

[Haffner 93] P. Haffner: Connectionist Speech Recognition with a Global MMI Algorithm. 3rd Europ. Conf. on Speech Communication and Technology (Eurospeech'93), Berlin, Germany, Sep. 1993.

[Hermansky & Ellis+ 00] H. Hermansky, D. W. Ellis, S. Sharma: Tandem connectionist feature extraction for conventional HMM systems. IEEE Int. Conf. on Acoustics, Speech and Signal Processing, pp. 1635-1638, Istanbul, Turkey, June 2000.

[Hinton & Osindero+ 06] G. E. Hinton, S. Osindero, Y. Teh: A fast learning algorithm for deep belief nets. Neural Computation, Vol. 18, No. 7, pp. 1527-1554, July 2006.

[Hochreiter & Schmidhuber 97] S. Hochreiter, J. Schmidhuber: Long short-term memory. Neural Computation, Vol. 9, No. 8, pp. 1735–1780, Nov. 1997.

[Klakow & Peters 02] D. Klakow, J. Peters: Testing the correlation of word error rate and perplexity. Speech Communication, pp. 19–28, 2002.

[Koehn & Och+ 03] P. Koehn, F. J. Och, D. Marcu: Statistical Phrase-Based Translation. HLT-NAACL 2003, pp. 48-54, Edmonton, Canada, May-June 2003.

[Le & Allauzen+ 12] 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.

[LeCun & Bengio+ 94] 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.

[Mikolov & Karafiat+ 10] 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.

[Nakamura & Shikano 89] M. Nakamura, K. Shikano: A Study of English Word Category Prediction Based on Neural Networks. ICASSP 89, p. 731-734, Glasgow, UK, May 1989.

[Ney 03] H. Ney: On the Relationship between Classification Error Bounds and Training Criteria in Statistical Pattern Recognition. First Iberian Conf. on Pattern Recognition and Image Analysis, Puerto de Andratx, Spain, Springer LNCS Vol. 2652, pp. 636-645, June 2003.

[Och & Ney 03] F. J. Och, H. Ney: A Systematic Comparison of Various Alignment Models. *Computational Linguistics,* Vol. 29, No. 1, pp. 19-51, March 2003.

[Och & Ney 04] F. J. Och, H. Ney: The Alignment Template Approach to Statistical Machine Translation. *Computational Linguistics*, Vol. 30, No. 4, pp. 417-449, Dec. 2004.

[Och & Tillmann+ 99] 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.

[Peddinti & Povey+ 15] V. Peddinti, D. Povey, S. Khudanpur: A time delay neural network architecture for efficient modeling of long temporal contexts. Interspeech, Dresden, Germany, Sep. 2015.

[Povey & Woodland 02] D. Povey, P.C. Woodland: Minimum phone error and I-smoothing for improved discriminative training. IEEE Int. Conf. on Acoustics, Speech and Signal Processing, pp. 105–108, Orlando, FL, May 2002.

[Printz & Olsen 02] H. Printz, P. A. Olsen: Theory and practice of acoustic confusability. Computer Speech and Language, pp. 131–164, Jan. 2002.

[Robinson 94] 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.

[Schlüter & Nussbaum+ 12] 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.

[Schlüter & Nussbaum-Thom+ 13] R. Schlüter, M. Nußbaum-Thom, E. Beck, T. Alkhouli, H. Ney: Novel Tight Classification Error Bounds under Mismatch Conditions based on f-Divergence. IEEE Information Theory Workshop, pp. 432–436, Sevilla, Spain, Sep. 2013.

[Schuster & Paliwal 97] M. Schuster, K. K. Paliwal: Bidirectional Recurrent Neural Networks. IEEE Trans. on Signal Processing, Vol. 45, No. 11, pp. 2673-2681, Nov. 1997.

[Schwenk 07] H. Schwenk: Continuous space language models. Computer Speech and Language, Vol. 21, No. 3, pp. 492–518, July 2007.

[Schwenk 12] 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.

[Schwenk & Costa-jussa+ 07] 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.

[Schwenk & Déchelotte+ 06] 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.

[Solla & Levin+ 88] S. A. Solla, E. Levin, M. Fleisher: Accelerated Learning in Layered Neural Networks. Complex Systems, Vol.2, pp. 625-639, 1988.

[Sundermeyer & Alkhouli+ 14] 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.

[Sundermeyer & Ney+ 15] M. Sundermeyer, H. Ney, R. Schlüter: From feedforward to recurrent LSTM neural networks for language modeling. IEEE/ACM Trans. on Audio, Speech, and Language Processing, Vol. 23, No. 3, pp. 13–25, March 2015.

[Sundermeyer & Schlüter+ 12] M. Sundermeyer, R. Schlüter, H. Ney: LSTM neural networks for language modeling. Interspeech, pp. 194–197, Portland, OR, USA, Sep. 2012.

[Utgoff & Stracuzzi 02] P. E. Utgoff, D. J. Stracuzzi: Many-layered learning. Neural Computation, Vol. 14, No. 10, pp. 2497-2539, Oct. 2002.

[Vaswani & Zhao+ 13] 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, USA, Oct. 2013.

[Vogel & Ney+ 96] 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.

[Waibel & Hanazawa+ 88] 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.

[Waibel & Hanazawa+ 89] A. Waibel, T. Hanazawa, G. Hinton, S. Kiyohiro, and K.J. Lang. “Phoneme recognition using time-delay neural networks,” *IEEE Trans. on Acoustics, Speech and Signal Processing*, Vol. 37, No. 3, pp. 328–339, 1989.

[Zens & Och+ 02] R. Zens, F. J. Och, H. Ney: Phrase-Based Statistical Machine Translation. 25th Annual German Conf. on AI, pp. 18–32, LNAI, Springer 2002.

END