Automatic Speech Recognition

Introduction to Deep Learning in ASR
+ Some examples

Alberto Abad

IST/INESC-ID Lisboa, Portugal

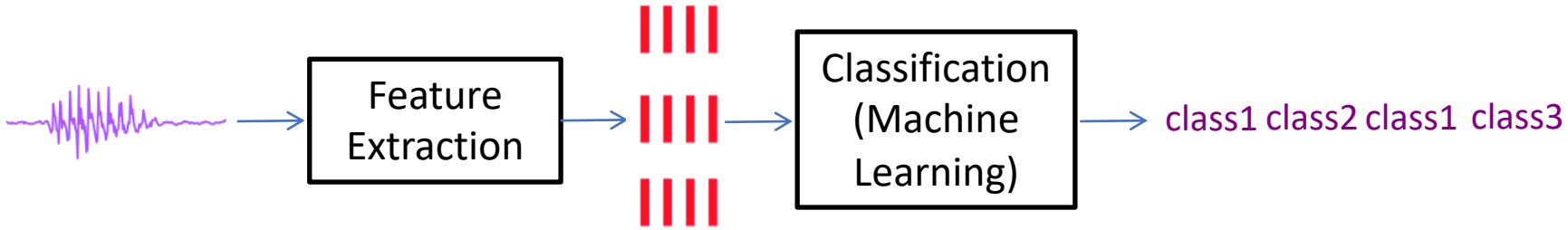
alberto.abad@tecnico.ulisboa.pt

Speech Processing - **IST** Lisboa, May, 2020

Introduction

Challenges from ML perspective (I)

- The common blocks of any speech pattern classification task are the front-end/feature extraction and the back-end/classification:



- From ML perspective, ASR is a very challenging problem due to the nature of the input and class label outputs
  - About the input → Time **sequence**
    - Very different length of the input wrt. output → Segmentation problem
    - Elasticity of the temporal dimension
    - Discriminative cues often distributed over a long temporal span
  - About the output → Output is a **sequence** of labels/words

sequence2sequence PROBLEM

Introduction

Challenges from ML perspective (II)

- Research in ASR has produced very significant outcomes during last decades (but it is still an open problem).
- Two main current trends to tackle the problem:

1. **Hierarchical modelling of speech**
   - Speech modelling problem is structured in sub-problems
   - This is the conventional approach until ~2012
   - Today still very relevant in certain tasks/conditions

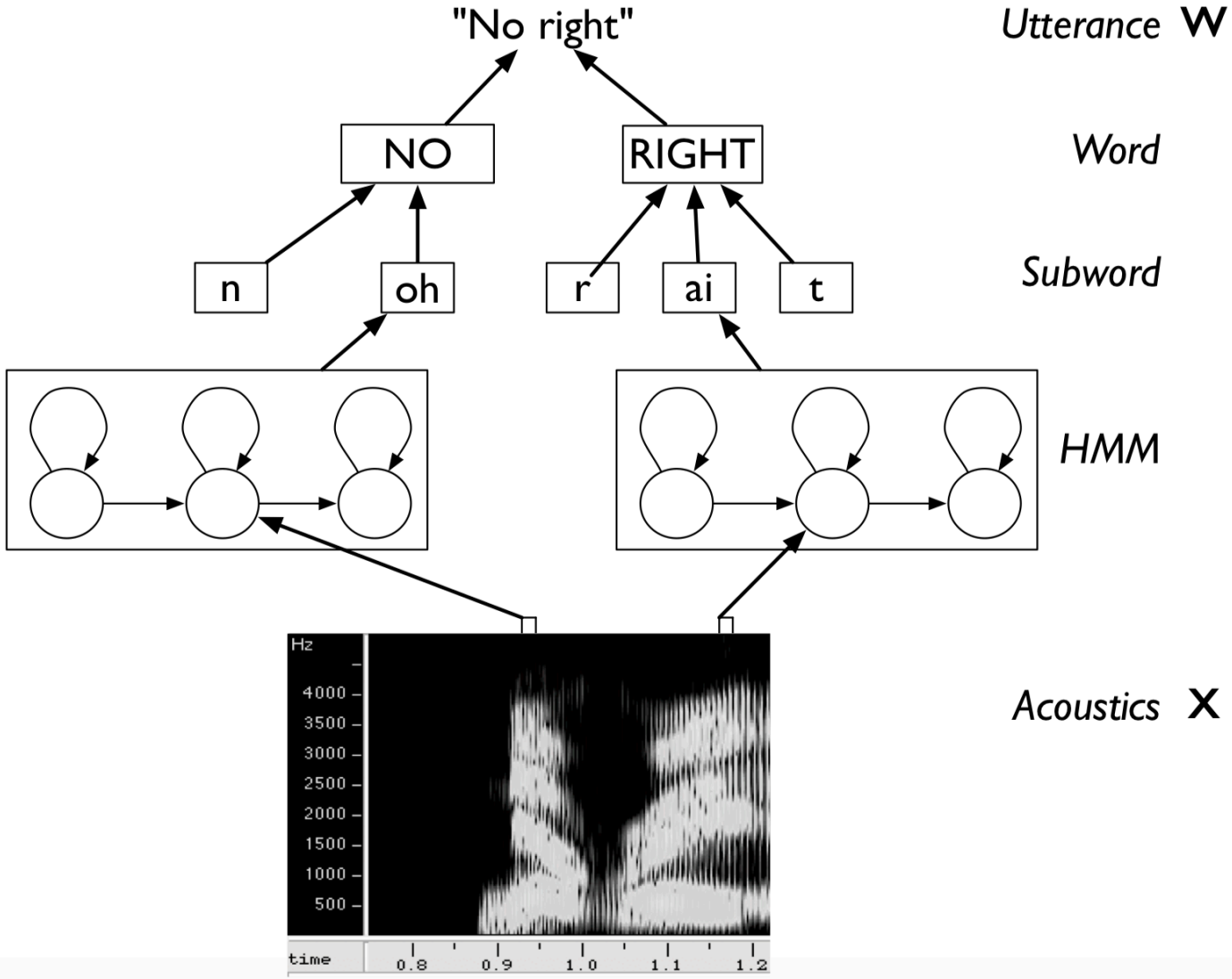
LAST WEEK → HMM/GMM
+
TODAY → HMM/ANN

2. **end2end**
   - Direct mapping from acoustics to words/characters
   - Different flavours from 2012 (CTC, encoder-decoder, etc.)
   - State of the art (in very large data)

NEXT THURSDAY!!!

Introduction

ASR: Hierarchical modeling of speech



Hybrid HMM/ANN
Automatic Speech Recognition

Hybrid HMM/ANN ASR

The output observation distribution

- The observation likelihood of conventional HMM/GMM approach:

$$b_j(x) = P(x|S = j) = \sum_{k=1}^{N} c_{jk} N[x|\mu_{jk}, \Sigma_{jk}]$$

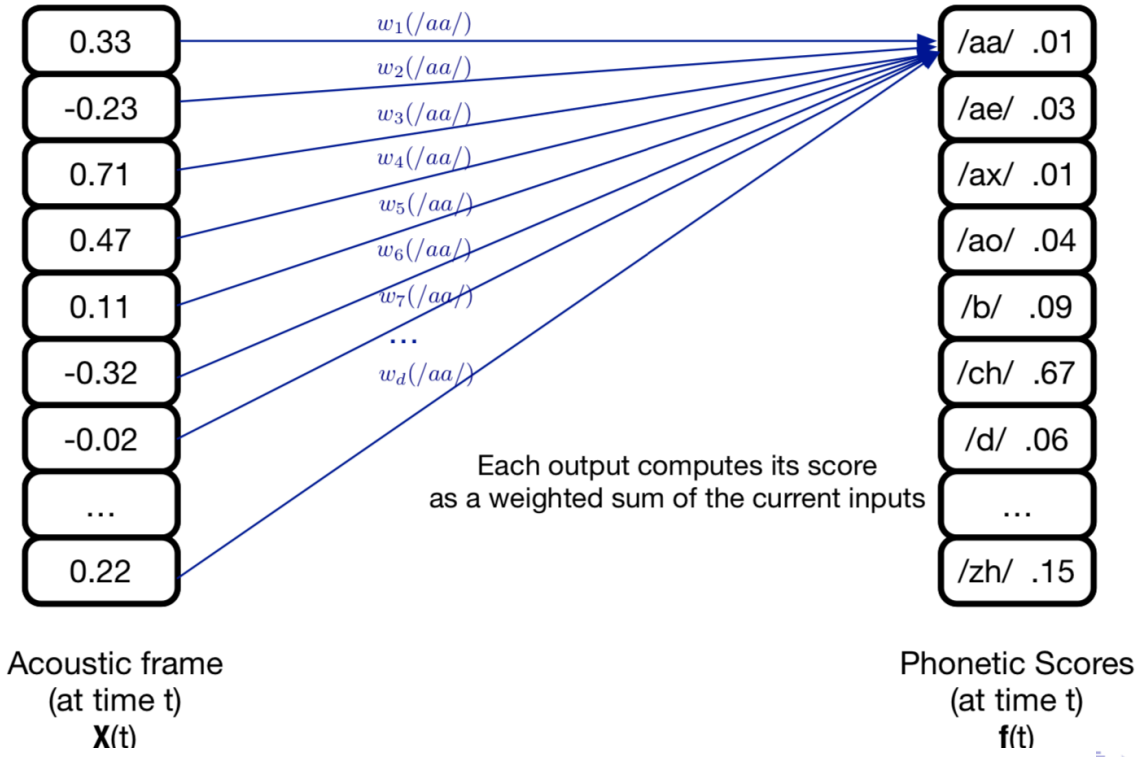
- It is now replaced by a (scaled) neural-network posterior:

$$b_j(x) \sim P(S = j|x)/P(S = j)$$

Hybrid HMM/ANN ASR

Simple neural network AM

- Input → Acoustic frame at time t, *x(t)*
- Output → Phonetic score, *P(S=j|x)*
- Network details:
  - Single (or multi) layer neural network with *softmax* output (probabilities)
  - The "phonetic score" of training data are 1s or 0s given by HMM/GMM forced alignment
  - Back-propagation with cross-entropy loss functions
  - Forward inference provides real-valued numbers corresponding to *P(S=j|x)*



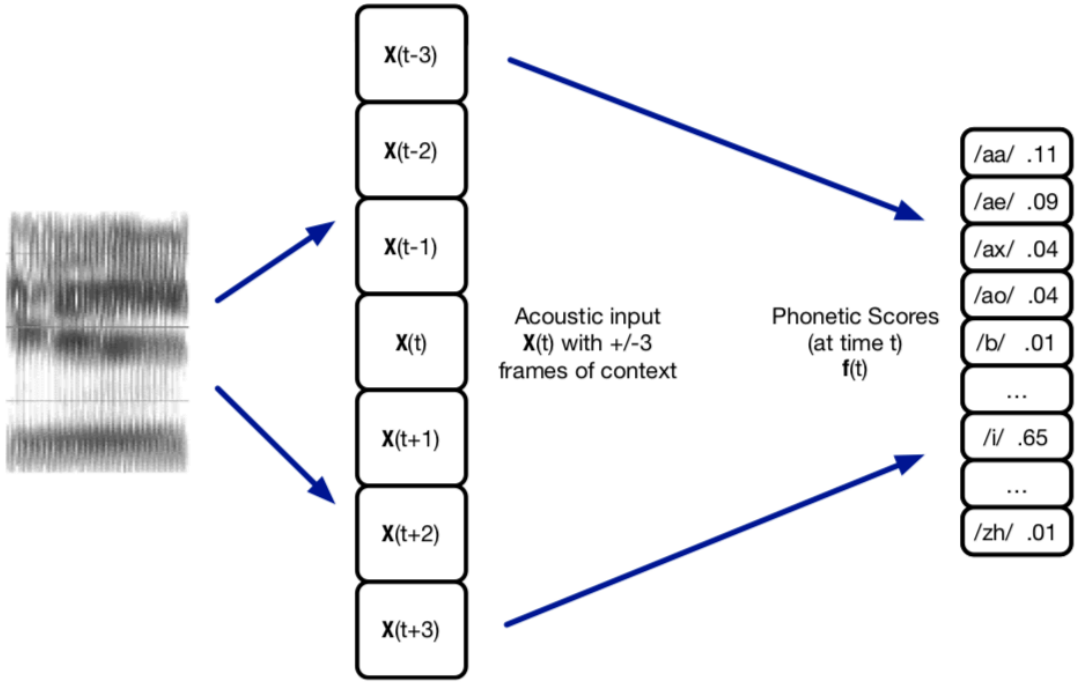
$$y_k = \frac{\exp(a_k)}{\sum_{j=1}^{K} \exp(a_j)}$$

$$a_k = \sum_{j=1}^{d} w_{kj} h_j + b_k$$

Hybrid HMM/ANN ASR

Simple neural network AM with acoustic context

- Input → Acoustic frames at time t +/- context, *x(t-context), …,x(t), .. x(t+context)*
- Output → Phonetic score, *P(S=j|x)*
- Network as previously



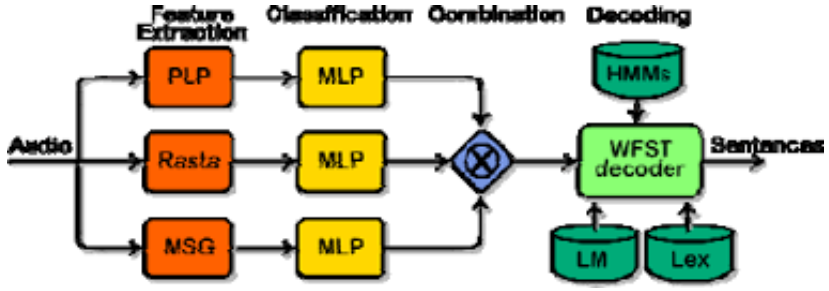
Hybrid HMM/ANN ASR

The connectionist approach (early 1990s)

- Monophone AM MLPs trained on multiple features showed:
  - Hybrid **monophone** MLP-based systems were superior than GMM-based counterparts (but worse than triphone)
  - NN can easily model correlated features:
    - Correlated feature vector components
    - Input context – multiple frames of data at input
  - NN more flexible than GMMs – GMMs inefficient for non-linear class boundaries
  - NNs can model multiple events and learn richer representations
  - NN posteriors are easy to combine

System	Nov92	si_dt_s6	si_dt_05.odd
HMM/GMM wint	8.11	10.39	12.40
HMM/GMM xword	6.86	9.52	10.48
ANN/HMM	9.73	13.13	14.37

Table 1: *WER results of HMM/GMM systems after [9] (word internal and cross-word) and of the baseline ANN/HMM system.*



Morgan and Bourlard (1995). Continuous speech recognition: Introduction to the hybrid HMM/connectionist approach, *IEEE Signal Processing Mag.*, 12(3):24-42

Hybrid HMM/ANN ASR

Extensions to the connectionist approach

- Model specific context-dependent units

System	Nov92	si_dt_s6	si_dt_05.odd	#Params
SS	9.77	13.13	14.37	650K (x3)
MS	8.74	12.59	13.56	700K (x3)
PT-20% (14)	8.52	11.53	12.52	710K (x3)
PT-40% (40)	8.66	11.23	11.91	730K (x3)
PT-50% (61)	7.79	10.27	11.88	740K (x3)
PT-60% (91)	8.28	10.09	11.64	760K (x3)
PT-70% (137)	7.98	9.22	11.22	800K (x3)
PT-80% (203)	7.57	9.61	10.95	840K (x3)

Table 1: *WER results of the baseline single-state (SS), multiple-state (MS), and phone transition (PT) ANN/HMM systems.*

System	Nov92	si_dt_s6	si_dt_05.odd	#Params
DD200	7.85	9.94	12.62	760K (x3)
DD300	7.77	10.99	11.35	830K (x3)
DD500	7.64	10.42	10.81	970K (x3)
DT200	7.75	11.36	12.08	760K (x3)
DT300	7.27	11.11	11.35	830K (x3)
DT500	7.38	11.20	11.95	970K (x3)

Table 2: *WER results of hybrid ANN/HMM systems for different number of context-dependent triphones with both data-driven (DD) and decision tree-based (DT) clustering.*

A. Abad and J. Neto (2008), Incorporating acoustical modelling of phone transitions in an hybrid ANN/HMM speech recognizer , *In* INTERSPEECH-2008, Brisbane (Australia), September 2008

A. Abad et al. (2010), Context Dependent Modelling Approaches for Hybrid Speech Recognizers , *In* Interspeech 2010, ISCA, Makuhari (Japan), September 2010

Hybrid HMM/ANN ASR

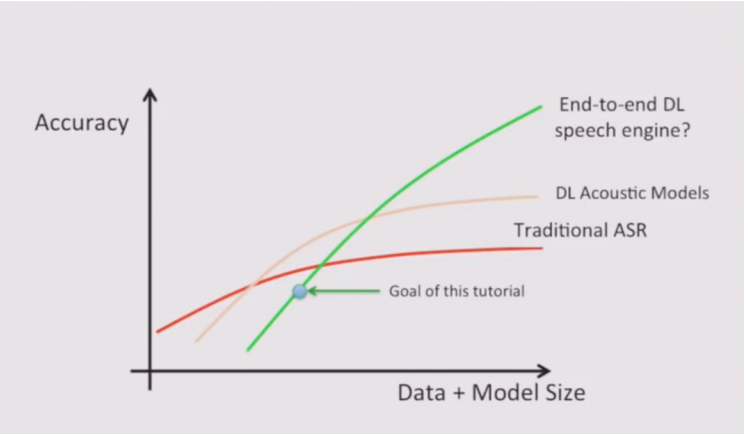
Disadvantages of NN until ~2012

- Context-independent (monophone) models
- Weak (speaker) adaptation algorithms
- NN systems less complex than GMMs (fewer parameters):
  - RNN < 100k parameters, MLP ~ 1M parameters
- Computationally expensive (still holds)
  - more difficult to parallelize training than GMM systems

Hybrid HMM/ANN ASR

What is different after 2012?

- DNNs proposed as AM for ASR:
  - **Deeper** networks, typical NNs AMs with 3-7 hidden layers:
    - This is partially possible to different advances in ML, including, regularization strategies, hidden unit non-linearity (ReLU vs tanh vs sigmoid), architectural choices
    - Initially, researchers thought pre-training was **THE TRICK**
      - Now, it is no longer relevant
  - **Wider** networks
    - Context-dependent HMM states or senones
  - **Computer/GPUs**
    - Permitted an increasing experiments
    - Scaling up data and parameters



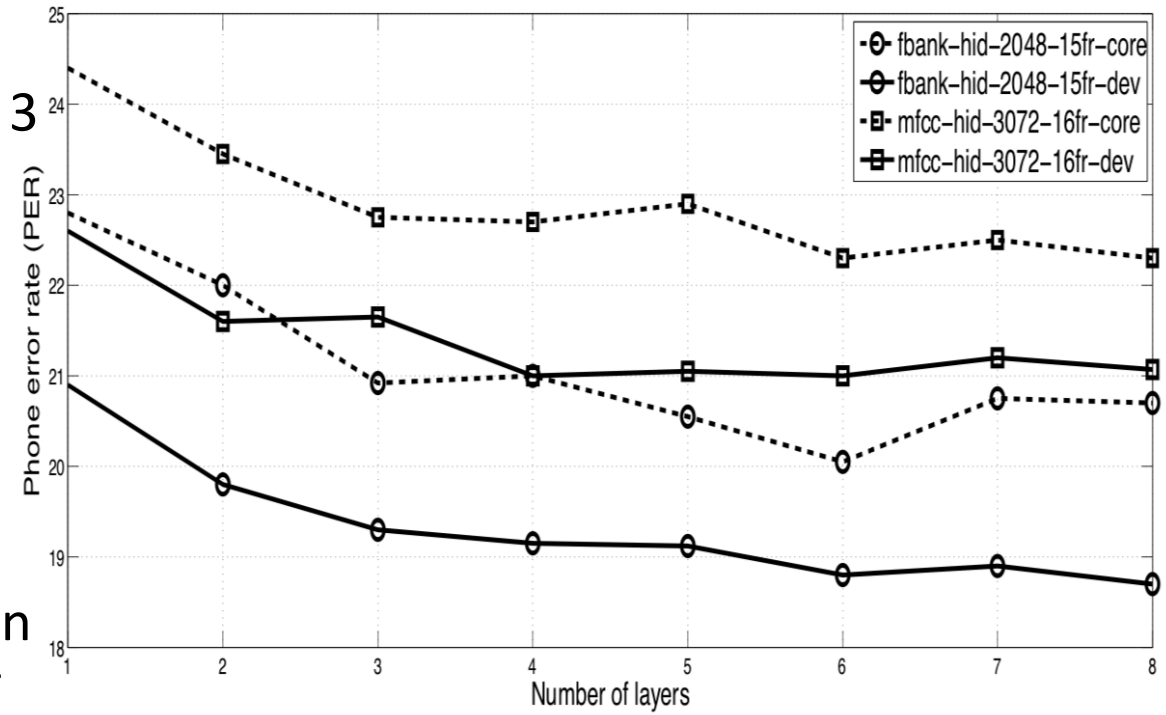
G. E. Dahl, et al (2012), Context-Dependent Pre-Trained Deep Neural Networks for Large-Vocabulary Speech Recognition, in *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 30-42, Jan. 2012.

Hinton, et al. (2012). Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups. Signal Processing Magazine, IEEE.

Hybrid HMM/ANN ASR

(Early) examples of CI-HMM-DNN in TIMIT

- DNN training targets (time state alignment) provided by a 'baseline' three state monophone HMM/GMM system (61 phones, 3 state HMMs)
  - DNN has 183 (61*3) outputs
- About hidden layers
  - exact sizes not highly critical
  - 3–8 hidden layers
    1024–3072 units per hidden layer
  - Multiple hidden layers always work better than one hidden layer
- Best systems have lower phone error rate than best HMM/GMM systems (using state-of-the-art techniques such as discriminative training, speaker adaptive training)



(Mohamed et al (2012))

A Mohamed et al (2012). "Understanding how deep belief networks perform acoustic modelling", Proc ICASSP-2012.

Hybrid HMM/ANN ASR

(Early) examples of CD-HMM-DNN in LVCSR

[TABLE 3] A COMPARISON OF THE PERCENTAGE WERs USING DNN-HMMs AND GMM-HMMs ON FIVE DIFFERENT LARGE VOCABULARY TASKS.

TASK	HOURS OF TRAINING DATA	DNN-HMM	GMM-HMM WITH SAME DATA	GMM-HMM WITH MORE DATA
SWITCHBOARD (TEST SET 1)	309	18.5	27.4	18.6 (2,000 H)
SWITCHBOARD (TEST SET 2)	309	16.1	23.6	17.1 (2,000 H)
ENGLISH BROADCAST NEWS	50	17.5	18.8	
BING VOICE SEARCH (SENTENCE ERROR RATES)	24	30.4	36.2	
GOOGLE VOICE INPUT	5,870	12.3		16.0 (>> 5,870 H)
YOUTUBE	1,400	47.6	52.3	

(Hinton et al (2012))

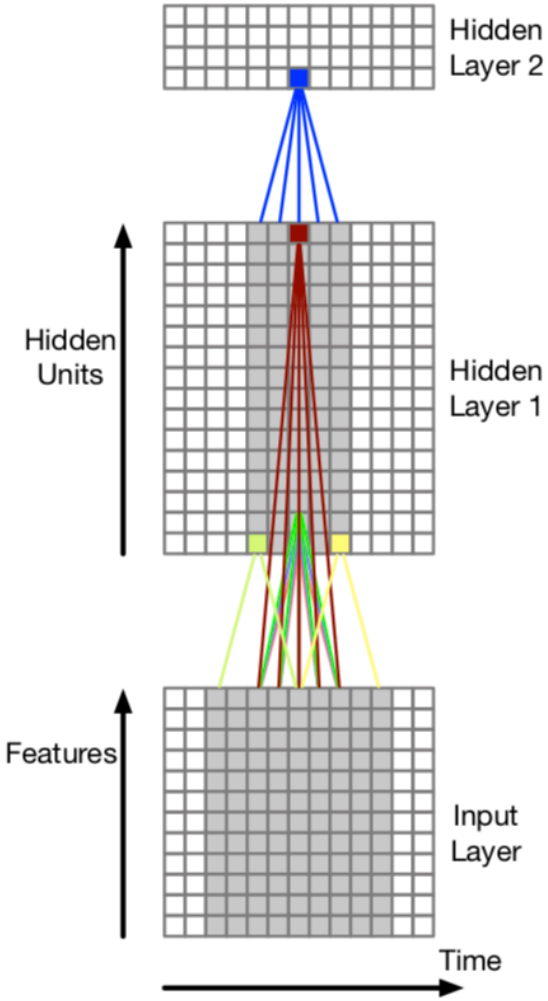
Hinton, et al. (2012). Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups. IEEE Signal Processing Magazine.

- Train a context-dependent HMM/GMM system, using a phonetic decision tree to determine the HMM tied states
  - Perform Viterbi alignment using the trained HMM/GMM and the training data
- Train a neural network using gradient descent to map the input speech features to a label representing a context-dependent tied HMM state
  - The size of the label set is thousands (number of context-dependent tied states)

Hybrid HMM/ANN ASR

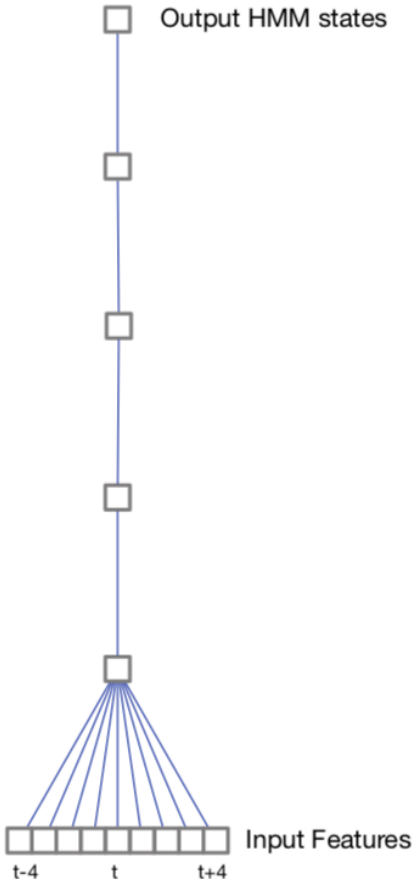
TDNNs – Richer acoustic context modelling

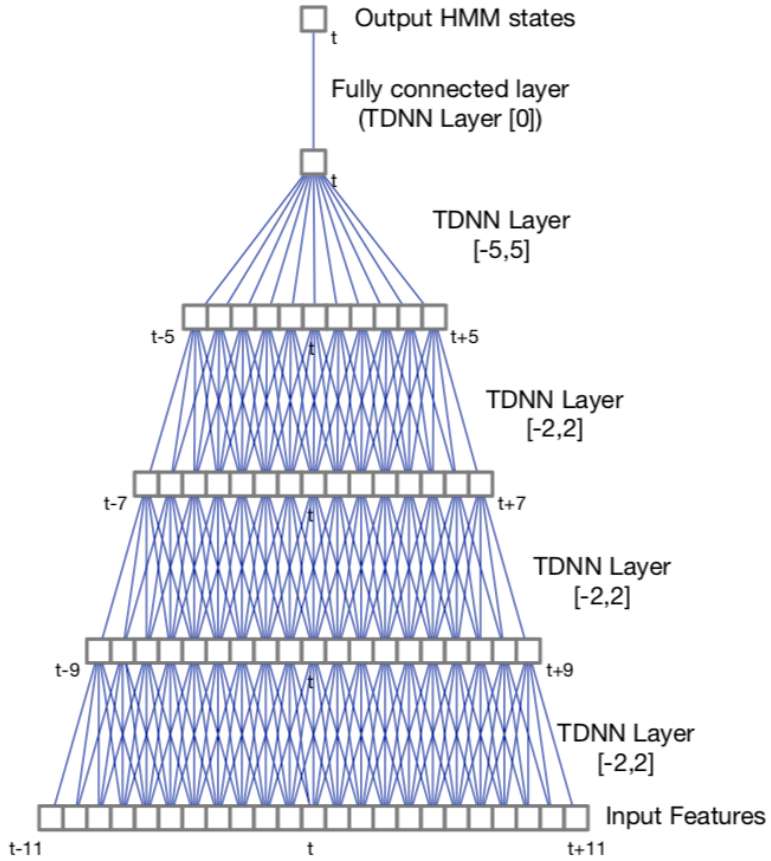
- Time-delay Neural Networks (TDNNs) model richer context:
  - each layer processes a context window from the previous layer
  - lower hidden layers learn from narrower contexts, higher hidden layers from wider acoustic contexts
  - higher hidden layers have a wider receptive field into the input

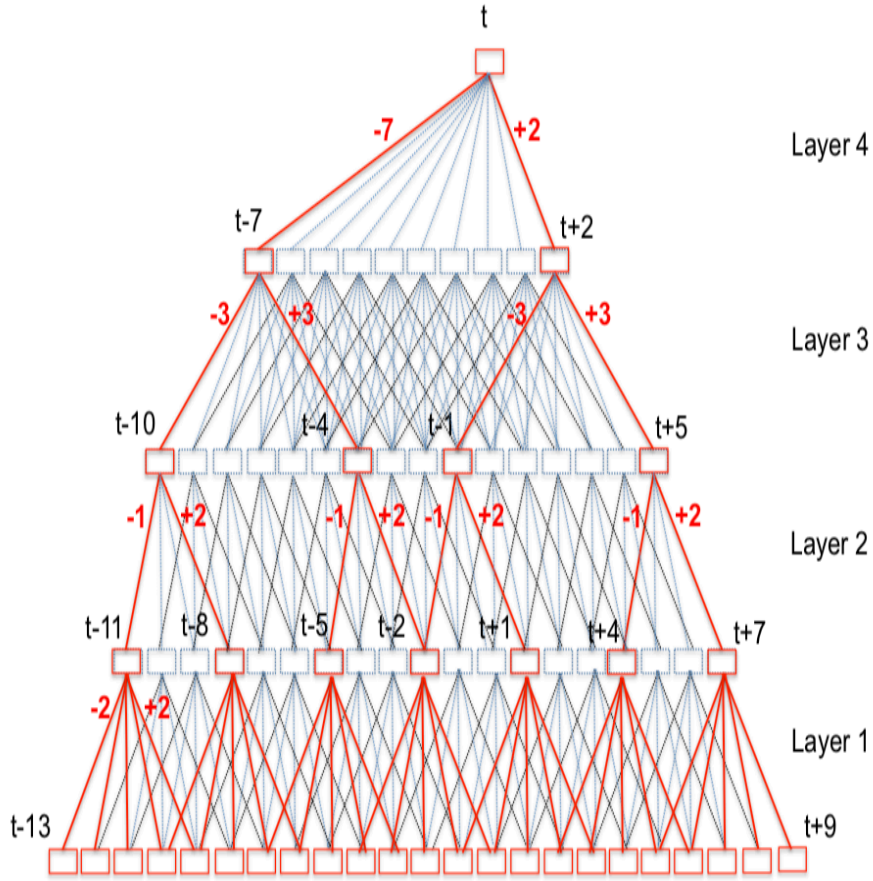


Hybrid HMM/ANN ASR

DNNs vs TDNNs vs sub-sampled TDNNs







Hybrid HMM/ANN ASR

DNNs vs TDNNs results (in SWB and other tasks)

Table 2: Performance comparison of DNN and TDNN with various temporal contexts

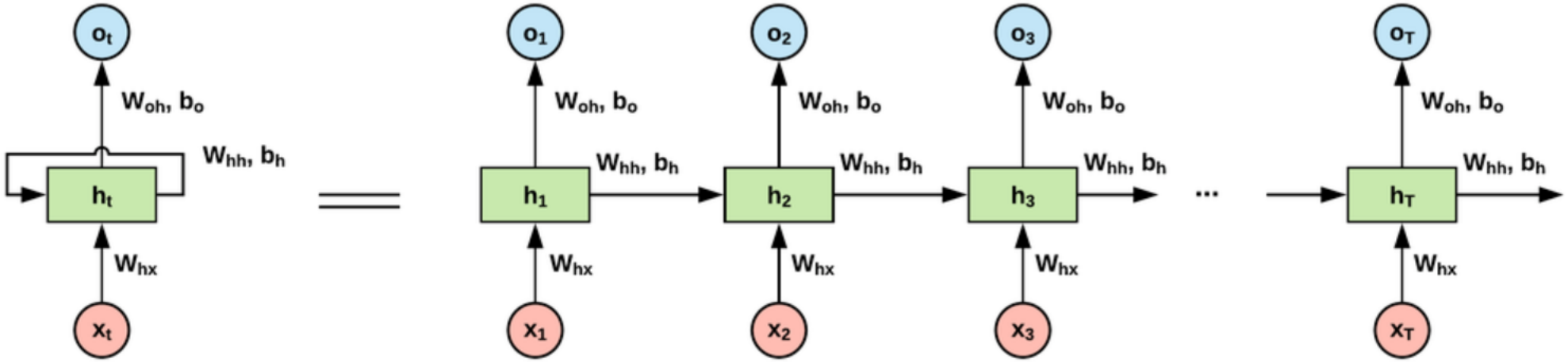
Model	Network Context	Layerwise Context					WER	
		1	2	3	4	5	Total	SWB
DNN-A	$[-7, 7]$	$[-7, 7]$	$\{0\}$	$\{0\}$	$\{0\}$	$\{0\}$	22.1	15.5
DNN-A_2	$[-7, 7]$	$[-7, 7]$	$\{0\}$	$\{0\}$	$\{0\}$	$\{0\}$	**21.6**	**15.1**
DNN-B	$[-13, 9]$	$[-13, 9]$	$\{0\}$	$\{0\}$	$\{0\}$	$\{0\}$	22.3	15.7
DNN-C	$[-16, 9]$	$[-16, 9]$	$\{0\}$	$\{0\}$	$\{0\}$	$\{0\}$	22.3	15.7
TDNN-A	$[-7, 7]$	$[-2, 2]$	$\{-2, 2\}$	$\{-3, 4\}$	$\{0\}$	$\{0\}$	21.2	14.6
TDNN-B	$[-9, 7]$	$[-2, 2]$	$\{-2, 2\}$	$\{-5, 3\}$	$\{0\}$	$\{0\}$	21.2	14.5
TDNN-C	$[-11, 7]$	$[-2, 2]$	$\{-1, 1\}$	$\{-2, 2\}$	$\{-6, 2\}$	$\{0\}$	20.9	14.2
TDNN-D	$[-13, 9]$	$[-2, 2]$	$\{-1, 2\}$	$\{-3, 4\}$	$\{-7, 2\}$	$\{0\}$	**20.8**	**14.0**
TDNN-E	$[-16, 9]$	$[-2, 2]$	$\{-2, 2\}$	$\{-5, 3\}$	$\{-7, 2\}$	$\{0\}$	20.9	14.2

Database	Size	WER		Rel. Change
		DNN	TDNN	
Res. Management	3h hrs	2.27	2.30	-1.3
Wall Street Journal	80 hrs	6.57	6.22	5.3
TedLIUM	118 hrs	19.3	17.9	7.2
Switchboard	300 hrs	15.5	14.0	9.6
Librispeech	960 hrs	5.19	4.83	6.9
Fisher English	1800 hrs	22.24	21.03	5.4

V Peddinti et al (2015). "A time delay neural network architecture for efficient modeling of long temporal contexts", Interspeech 2015.

Hybrid HMM/ANN ASR

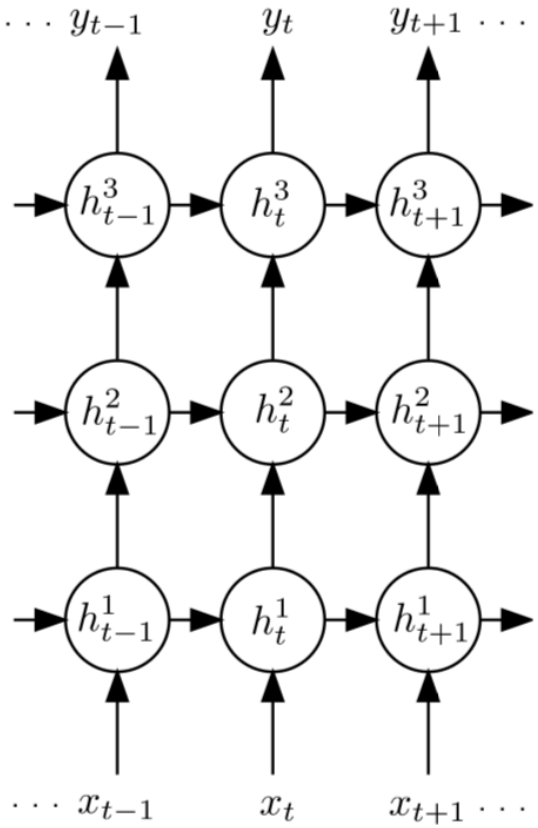
RNNs – Richer acoustic context modelling



- Recurrent Neural Networks (RNNs) allow to model richer context:
  - Hidden units at time t take input from their value at time t − 1
  - can be seen as a sequence of T inputs as a T -layer network with shared weights:
    - Train with backprop through time (BTT)
  - Recurrent hidden units are state units: can keep information through time
    - State units as memory – remember things for (potentially) an infinite time
    - State units as information compression – compress the history (sequence observed up until now) into a state representation

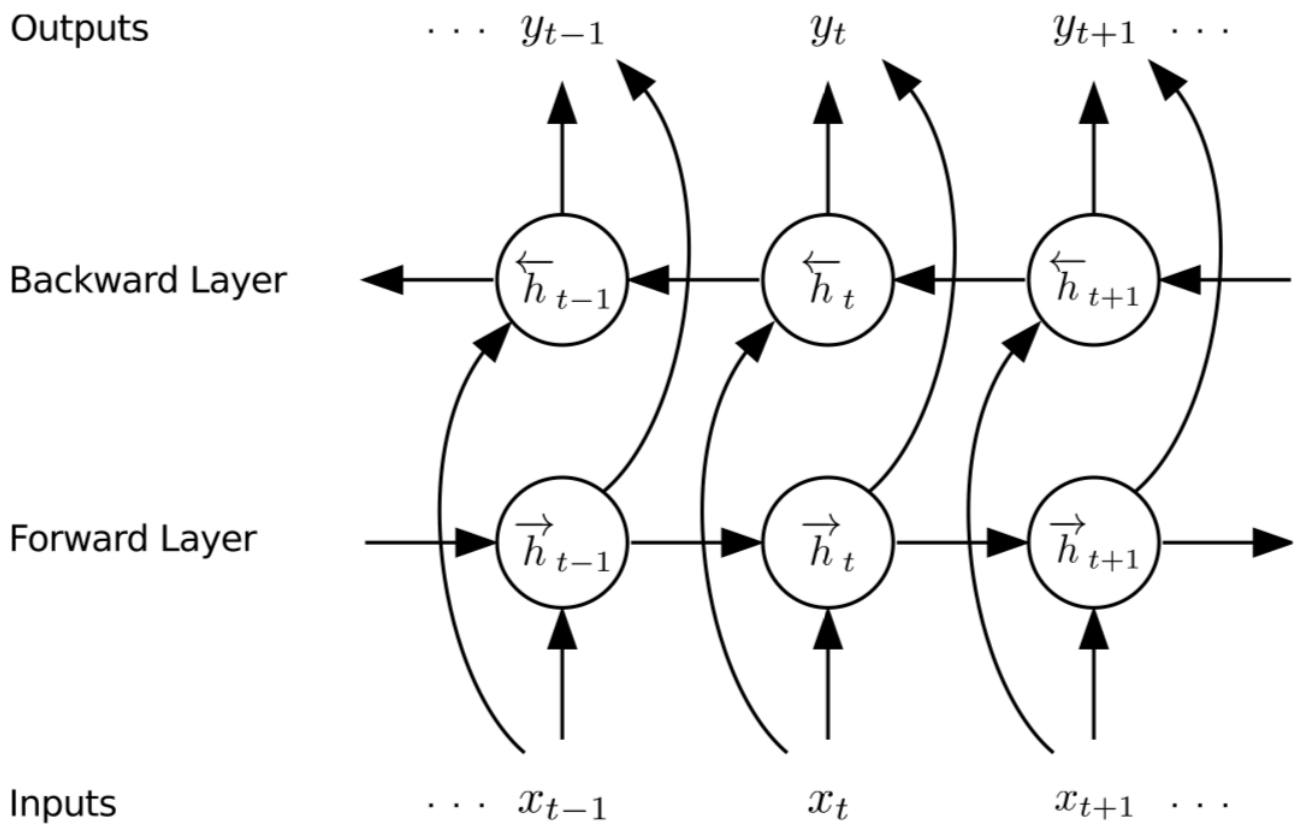
Hybrid HMM/ANN ASR

RNNs – Deep RNNs



Hybrid HMM/ANN ASR

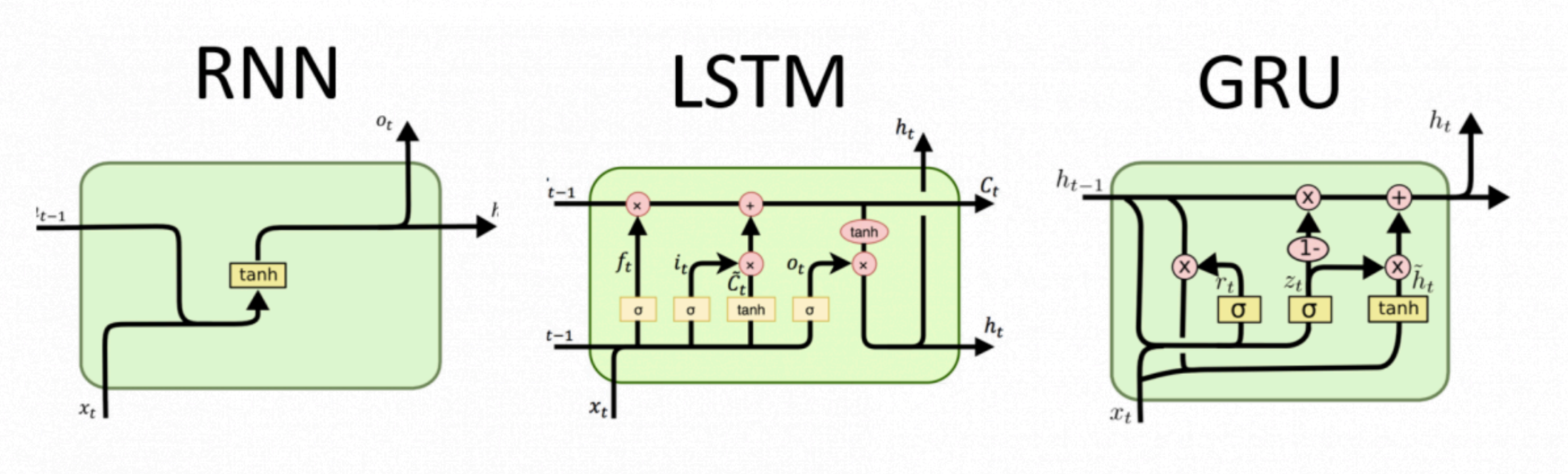
RNNs – Bidirectional RNNs



Hybrid HMM/ANN ASR

RNNs vs LSTMs vs GRUs

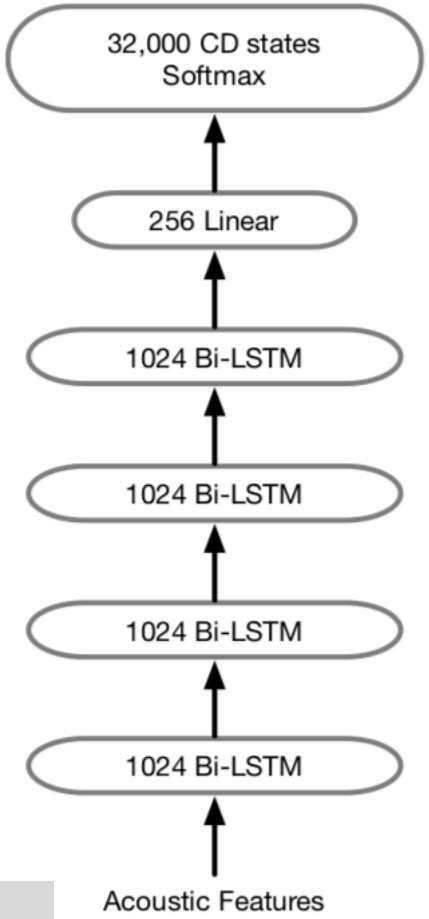
- Long-short term memory and Gated-recurrent units are special case of recurrent neural network
- Specifically designed to avoid “forgetting” in long input sequence problems (such as speech)



Hybrid HMM/ANN ASR

Deep Bidirectional LSTMs example

- LSTM with 4-6 bidirectional layers with:
  - 1024 cells/layer (512 each direction)
  - 256 unit linear bottleneck layer
  - 32k context-dependent state outputs
- Input features
  - 40-dimension linearly transformed MFCCs (plus ivector)
  - 64-dimension log mel filter bank features (plus first and second derivatives)
- Training: 14 passes frame-level cross-entropy training, 1 pass sequence training



Saon et al (2017), "English Conversational Telephone Speech Recognition by Humans and Machines", Interspeech-2017.

Hybrid HMM/ANN ASR

Deep Bidirectional LSTMs example

Network Architecture	Test Set WER/% Switchboard	CallHome
GMM (ML)	21.2	36.4
GMM (BMMI)	18.6	33.0
DNN (7x2048) / CE	14.2	25.7
DNN (7x2048) / MMI	12.9	24.6
TDNN (6x1024) / CE	12.5	
TDNN (6x576) / LF-MMI	9.2	17.3
LSTM (4x1024)	8.0	14.3
LSTM (6x1024)	7.7	14.0
LSTM-6 + feat fusion	7.2	12.7

GMM and DNN results - Vesely et al (2013); TDNN-CE results - Peddinti et al (2015); TDNN/LF-MMI results - Povey et al (2016); LSTM results - Saon et al (2017)

Combining models, and with multiple RNN language models, WER reduced to 5.5/10.3% (Saon et al, 2017)

Vesely et al. (2013), "Sequence-discriminative training of deep neural networks", in Proc. Interspeech, 2013.

Hybrid HMM/ANN ASR

Additional topics for discussion

- Data augmentation
- Multi-lingual, BNF features, etc.
- Multi-stage decoding/re-scoring
- Speaker adaptation (*)
- Sequence-training (*)
- My recent work on domain adaptation (*Alberto)
- Atypical speech in ASR (*Thomas)

Hybrid HMM/ANN ASR

Speaker adaptation: SAT based features

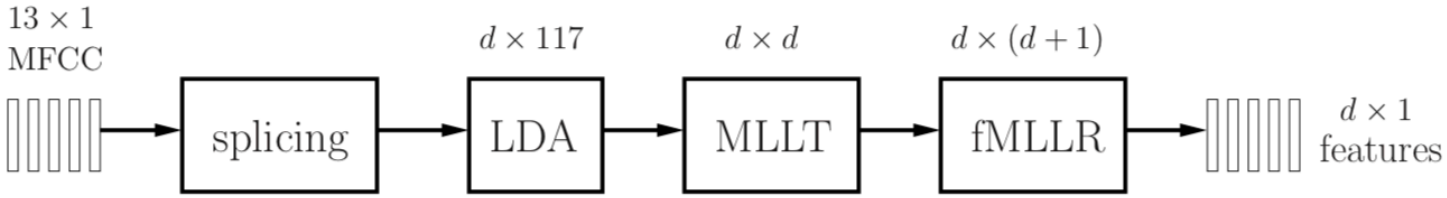


Figure 1: Generation of our baseline/Type I features

Table 1: WER (%) with GMM system using baseline features. The results are shown on Hub5'00-SWB and Hub5'00 (shown in brackets) test sets.

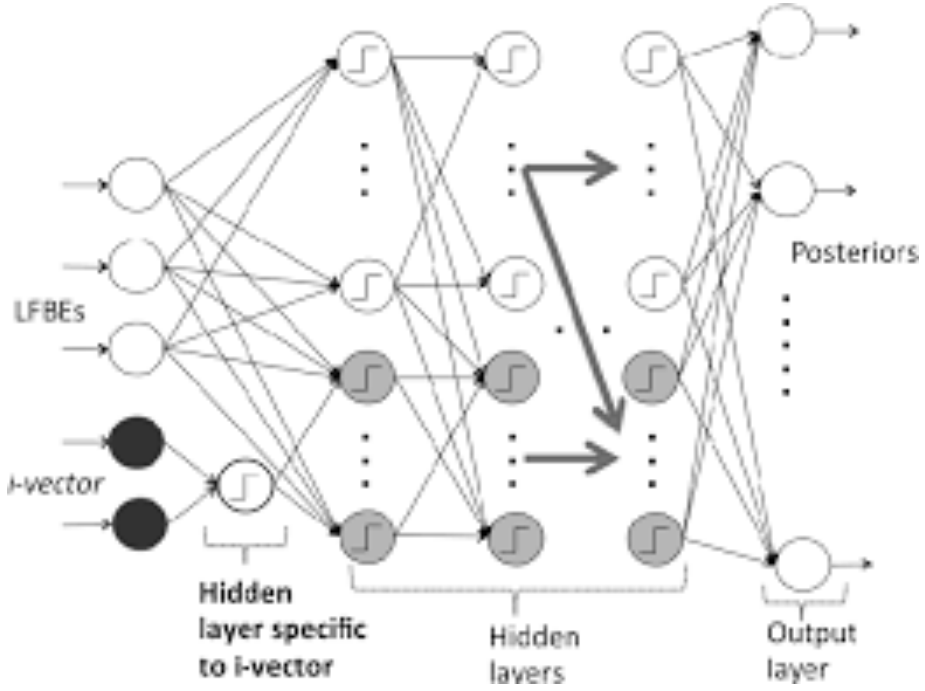
Type of feature	WER (%)
LDA+MLLT (no adaptation)	34.6 (42.5)
+fMLLR in test time	26.9 (34.4)
+fMLLR train/test (SAT)	25.6 (32.7)

Table 4: WER (%) with DNN using baseline/Type I features

d	LDA+MLLT (un-adapted)	+fMLLR test	+fMLLR train/test (SAT)
40	25.3 (32.6)	22.9 (29.4)	22.0 (28.4)
60	23.4 (30.6)	21.6 (28.0)	19.7 (26.5)
80	23.4 (30.1)	21.5 (27.7)	**19.5 (26.1)**
100	22.9 (29.9)	21.2 (27.4)	19.8 (26.2)
117	23.4 (30.4)	21.7 (28.0)	20.0 (26.4)

Hybrid HMM/ANN ASR

Speaker adaptation: Speaker coding/i-vectors



Model	Training	Hub5'00 SWB	RT'03 FSH	RT'03 SWB
DNN-SI	x-entropy	16.1%	18.9%	29.0%
DNN-SI	sequence	14.1%	16.9%	26.5%
DNN-SI+ivecs	x-entropy	13.9%	16.7%	25.8%
DNN-SI+ivecs	sequence	12.4%	15.0%	24.0%
DNN-SA	x-entropy	14.1%	16.6%	25.2%
DNN-SA	sequence	12.5%	15.1%	23.7%
DNN-SA+ivecs	x-entropy	13.2%	15.5%	23.7%
DNN-SA+ivecs	sequence	11.9%	14.1%	22.3%

TABLE I
COMPARISON OF WORD ERROR RATES FOR VARIOUS DNNS ON HUB5'00 AND RT'03 WITHOUT AND WITH HESSIAN-FREE SEQUENCE TRAINING.

G. Saon et al. (2013), "Speaker adaptation of neural network acoustic models using i-vectors," in Proc. ASRU 2013

Hybrid HMM/ANN ASR

Sequence training

- In conventional HMM/GMM systems as **alternative discriminative sequence training criteria** to conventional ML:
  - Scalable minimum Bayes risk (sMBR); Minimum phone error (MPE); Maximum mutual information (MMI)

- Similar approaches for DNN → Alternative loss functions to CE that take into account **sequence information**:
  - First approaches, first CE training followed by sequence discriminative training. Need decoding lattices of training data (inefficient).
  - Recent approaches, such as **Lattice-free MMI** (aka in Kaldi recipes as CHAIN model) train DNNs directly using a sequence discriminative criteria:
    - Actually, CE is used as a regularization step
    - No need for previous lattice decoding
    - Introduce several tricks, including HMMs topology modifications and frame rate decimation

Hybrid HMM/ANN ASR

Sequence training: Lattice-free MMI

- LF-MMI introduce remarkable improvements:
  - In training/decoding times
  - WER performance

Objective	Model (size)	WER (%)
CE	TDNN-A (16.6M)	12.5
CE $\rightarrow$ sMBR	TDNN-A (16.6M)	11.4
LF-MMI	TDNN-A (9.8M)	10.7
	TDNN-B (9.9M)	10.4
	TDNN-C (11.2M)	10.2
LF-MMI $\rightarrow$ sMBR	TDNN-C (11.2M)	10.0

D. Povey, et al. (2016), "Purely sequence-trained neural networks for ASR based on lattice-free MMI" in Proc. Interspeech, 2016.

Hybrid HMM/ANN ASR

Domain adaptation for low-resource ASR

- **Goal:** Transfer specific channel/style conditions learnt in a well-resourced (WR) language to a low-resourced (LR) language for which training data is not available
- **How?**
  - Train multi-lingual/multi-task AM with WR+LR data in a common channel/style (ie. CTS).
  - Adapt network using new channel/style WR data (ie. BN):
    - Adapt only (at most) up to the last common layer, so the last language specific layers an unchanged.
  - Transfer adapted first layer weights and concatenate with LR last layers.
- Related with transfer learning, model adaptation, low resource ASR, multi-lingual learning, etc.

Abad et al. (2020), "Cross lingual transfer learning for zero-resource domain adaptation", Proc. ICASSP 2020

Hybrid HMM/ANN ASR

Domain adaptation for low-resource ASR

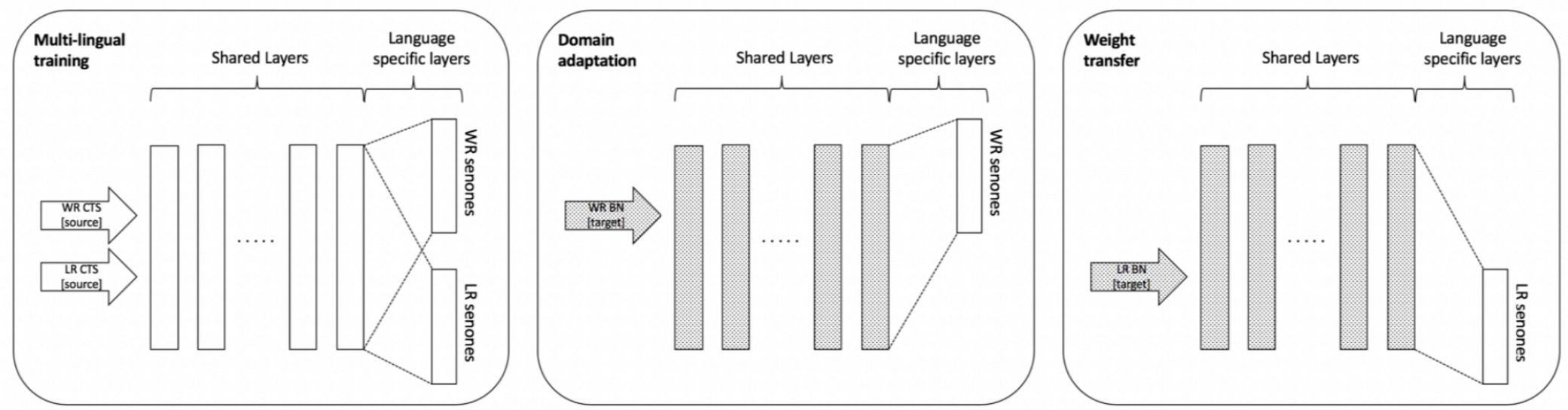


Figure 1: *The basic steps of the proposed cross-lingual domain adaptation scheme: 1) multi-lingual training; 2) adaptation of the shared parameters using WR data in the target domain; and 3) weight transfer the domain adapted shared layers to the original LR final language-dependent layers.*

Hybrid HMM/ANN ASR

Domain adaptation for low-resource ASR

	Test condition			
	WR language		LR language	
	CTS source	BN target	CTS source	BN target
mono-ling BN AM	---	11.8	---	19.2*
mono-ling CTS AM	22.6	19.6	32.3	40.0
multi-ling CTS AM	23.6	19.2	32.6	**32.9**

- CTS (Fisher) is the source condition and BN (hub4) is the target condition
- Spanish is the LR language and English the WR language
- Experimental set-up:
  - TDNN hires + pitch, no LF-MMI, no ivecs, all downsampled to 8kHz
  - Use of matched LMs (CTS/BN test data is decoded with CTS/BN LM)

Hybrid HMM/ANN ASR

Domain adaptation for low-resource ASR

	WR language	LR language
	BN target	BN target
Upper bound	11.8	19.2
mono-ling CTS AM	19.6	40.0
multi-ling CTS AM	19.2	32.9
proposed CL adapt AM	14.5	**28.4**

- From 40.0% to 32.9% thanks to multilang and from 32.9% to 28.4 to nnet adapt & transfer learning → NO USE OF ANY ADDITIONAL LR TRAINING DATA!!!

Domain adaptation for low-resource ASR: Experiments with low-resourced languages

- Use BABEL training set:
  - Exact same architecture as previous experiments
- Eval on BABEL dev and Material *analysis_** test sets:
  - CSTR MATERIAL LM → Trained on *webnews*

	Tagalog			Lithuanian		
	BN	TB	avg	BN	TB	avg
mono-ling CTS AM	53.2	58.7	57.3	45.6	43.0	44.0
multi-ling CTS AM	46.5	52.2	50.7	38.2	36.5	37.1
proposed CL adapt AM	41.9	48.5	**46.8**	31.6	32.1	**31.9**

- Same network architecture, training, decoding recipes and adaptation configuration (3 first hidden shared layers adapted for 1 epoch)
- Remarkable improvements in any of the two wide-band sub-domains:
  - BN: relative WER improvements of 21.2% for Tagalog and 30.7% for Lithuanian;
  - TB: 17.4% for Tagalog and 25.3% for Lithuanian.
- Overall, average relative WER improvement of 18.3% and 27.5% for the Tagalog and Lithuanian.

Summary

- LVCSR has witnessed great improvements since 2012 due to the positive impact of deep learning
- First generation deep learning based ASR systems replace the AM of a hierarchical/statistical conventional system by a DNN:
  - Better leverage of data
  - Better context modeling
  - Better accuracies (improving by a large margin long-standing SOA)
- Other key components contributed to improvements in HMM/DNN:
  - Architectural and ML choices
  - Data augmentation techniques
  - Side-speaker information for SAT
  - Sequence discriminative training
  - Transfer learning methods

Proposed exercise (for those interested) using KALDI toolkit

- Install and compile KALDI in a machine with GPUs
- Identify in one of the KALDI recipes (for instance, *librispeech*) the different modules and techniques introduced in this seminar:
  - Understand the role of each script and technique at an high-level
- Run one of these recipes (for instance, *librispeech* or *minilibirispeech*)

References (I)

Gales, M.J.F. & Young, Steve (2007) The Application of Hidden Markov Models in Speech Recognition. Foundations and Trends in Signal Processing. 1. 195-304.

Morgan and Bourlard (1995). Continuous speech recognition: Introduction to the hybrid HMM/connectionist approach, *IEEE Signal Processing Mag.*, 12(3):24-42

A. Abad and J. Neto (2008), Incorporating acoustical modelling of phone transitions in an hybrid ANN/HMM speech recognizer , *In* INTERSPEECH-2008, Brisbane (Australia), September 2008

A. Abad et al. (2010), Context Dependent Modelling Approaches for Hybrid Speech Recognizers , *In* Interspeech 2010, ISCA, Makuhari (Japan), September 2010

G. E. Dahl, et al (2012), Context-Dependent Pre-Trained Deep Neural Networks for Large-Vocabulary Speech Recognition, in *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 30-42, Jan. 2012.

A Mohamed et al (2012). “Understanding how deep belief networks perform acoustic modelling”, Proc ICASSP-2012.

References (II)

Hinton, et al. (2012). Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups. IEEE Signal Processing Magazine.

V Peddinti et al (2015). “A time delay neural network architecture for efficient modeling of long temporal contexts”, Interspeech 2015.

Saon et al (2017), “English Conversational Telephone Speech Recognition by Humans and Machines”, Interspeech-2017.

Vesely et al. (2013), “Sequence-discriminative training of deep neural networks”, in Proc. Interspeech, 2013.

G. Saon et al. (2013), “Speaker adaptation of neural network acoustic models using i-vectors,” in Proc. ASRU 2013

D. Povey, et al. (2016), “Purely sequence-trained neural networks for ASR based on lattice-free MMI” in Proc. Interspeech, 2016.

Abad et al. (2020), “Cross lingual transfer learning for zero-resource domain adaptation”, in Proc. ICASSP, 2020.