Aalto University

Deep learning for automatic speech recognition

Mikko Kurimo

Department for Signal Processing and Acoustics

Aalto University

Mikko Kurimo

Associate professor in **speech and language processing**

Background from machine learning algorithms and pattern recognition systems

PhD 1997 at TKK on **speech recognition with neural networks**

Research experience in several top speech groups:

- Research Centers: IDIAP (CH), SRI (USA), ICSI (USA)
- Universities: Edinburgh, Cambridge, Colorado, Nagoya

Head of Aalto **speech recognition research group +** several national and European speech and language projects

Research topics:

- Speech recognition, language modeling, speaker adaptation, speech synthesis, translation, information retrieval from audio and video

Contents of this talk

1. Applications of Automatic Speech Recognition (ASR)
2. Building blocks in ASR systems
3. Deep neural networks (DNN) for acoustic models (AM)
4. Deep neural networks (DNN) for language models (LM)

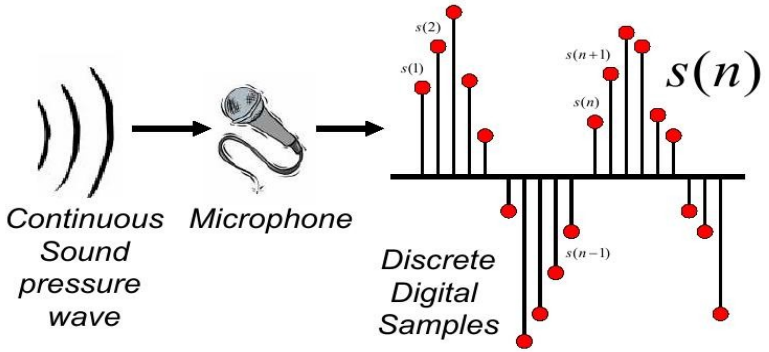
Using Automatic Speech Recognition (ASR)

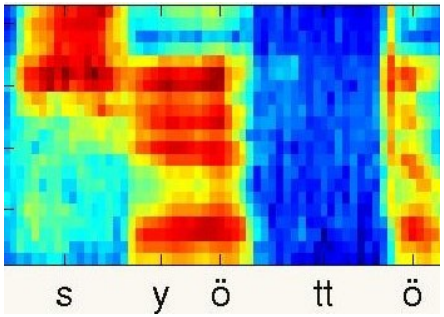
Mapping human speech to text or commands.

Has quickly become popular via voice search and virtual assistants in phones (Google, Siri etc).

Other applications: subtitling or indexing video recordings and streams, robots, toys, games, dictation, speech translation, disabled users, language learning and other education

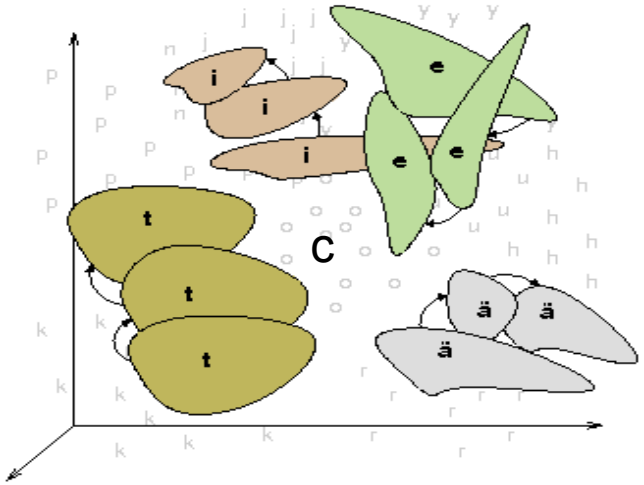
Building blocks in ASR systems

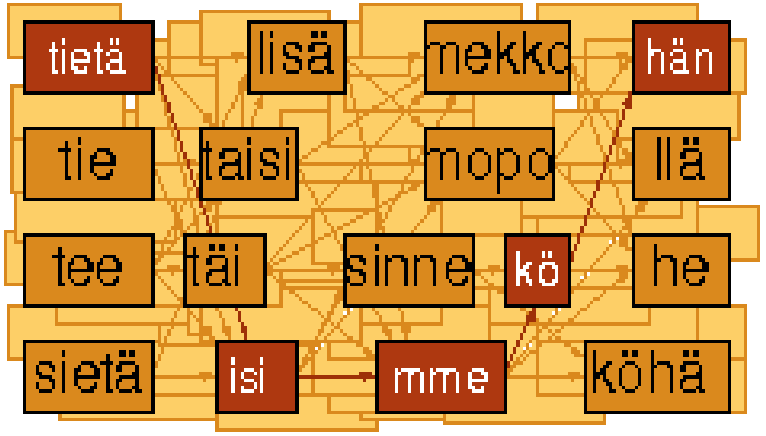




Language models

Speech → 10 ms features → Acoustic models → Decoder → Text





ASR performance depends on:

- Training and development data:
  - Quantity and suitability
- Recording and noise:
  - Microphone and distance
- Speakers and speaking styles:
  - Speaker changes
  - Clarity and style
- Language styles:
  - Grammatical vs colloquial
  - Planned vs spontaneous
  - Non-standard vocabulary

Why deep learning is needed in ASR?

1. **Acoustic** models (AM)

   - complicated density functions in time and frequency

   - variability between speakers

   - variability between styles

2. **Language** models (LM)

   - complicated dependencies at various levels: syntax, semantics, pragmatics

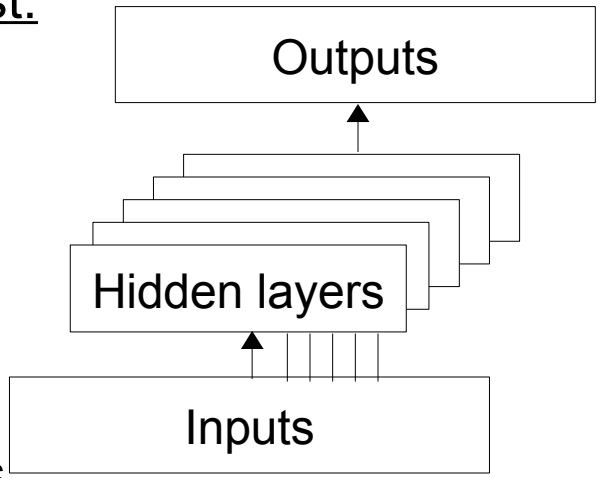
   - long-range dependencies

   - spontaneous speech is hard

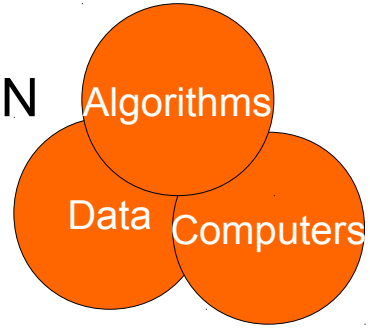
Analysis of DNNs in acoustic models (AM)

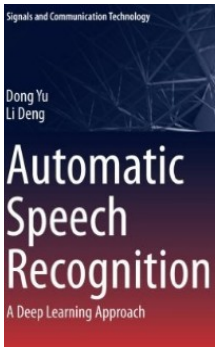
3 key ideas in DNNs that improve ASR most:

1. Processing in many hierarchical layers
2. Input from many frames
3. Output for context-dependent phones



Other significant improvements: speedups, pre-training, sequence discriminative training, multitask learning, various NN architectures (CNN, RNN, LSTM, Highways, Attention)





See: D.Yu, L.Deng. Automatic Speech Recognition A Deep Learning Approach. Springer 2015.

Unsolved research problems for DNN AM

1. Adaptation into new situations with little data (1,2,5)
2. Far field microphones, noisy and reverberant conditions (3,4)
3. Accented and dialect speech (5,6)
4. Spontaneous, non-fluent, and emotional speech (1,6)

(1) M.Kurimo, S.Enarvi, O.Tilk, M.Varjokallio, A.Mansikkaniemi, T.Alumäe. Modeling under-resourced languages for speech recognition. Language Resources and Evaluation, pp.1—27, 2016.

(2) P.Smit, J.Leinonen, K.Jokinen, M.Kurimo. Automatic Speech Recognition for Northern Sámi with comparison to other Uralic Languages. Proc. IWCLUL 2016.

(3) H.Kallasjoki. Feature Enhancement and Uncertainty Estimation for Recognition of Noisy and Reverberant Speech. PhD thesis. Aalto University, 2016.

(4) U.Remes. Statistical Methods for Incomplete Speech Data. PhD thesis. Aalto University, 2016.

(5) P.Smit, M.Kurimo. Using stacked transformations for recognizing foreign accented speech. Proc. ICASSP 2011.

(6) R.Karhila, A.Rouhe, P.Smit, A.Mansikkaniemi, H.Kallio, E.Lindroos, R.Hildén, M.Vainio, M.Kurimo. Digitala: An augmented test and review process prototype for high-stakes spoken foreign language examination. In Show and Tell at Interspeech 2016 .

Deep learning in language models (LM)

Steps taken from conventional LMs to DNNs:

1. Smoothed and pruned N-gram LMs (e.g. modified Kneser-Ney, Varigrams) (1,2)
2. Continuous space models using N-gram features (e.g. Maximum Entropy LMs) (3,4)
3. Neural Network LMs with input on different time scales (e.g. Recurrent NNs, Long Short Term Memory) (5)

(1) V.Siivola, M.Creutz, M.Kurimo. Morfessor and VariKN machine learning tools for speech and language technology. Proc. Interspeech 2007.

(2) T.Hirsimäki, J.Pylkkönen, M.Kurimo. Importance of high-order n-gram models in morph-based speech recognition. IEEE Trans. on Audio, Speech and Language Processing, 17(4), 2009.

(3) V.Siivola, A.Honkela. A state-space method for language modeling. Proc. ASRU 2003.

(4) T.Alumäe, M.Kurimo. Domain adaptation of maximum entropy language models. Proc. ACL 2010.

(5) S.Enarvi, M.Kurimo. TheanoLM - An Extensible Toolkit for Neural Network Language Modeling. Proc. Interspeech 2016.

Research problems in DNN LM

1. **Input & Output**: What are the basic modeling units (words, morphemes, letters) and their most effective and scalable embeddings (1,2)
2. **Network structure**: How to take into account both short-term (syntax, n-grams) and long-term (topics, referencing) dependences (3)

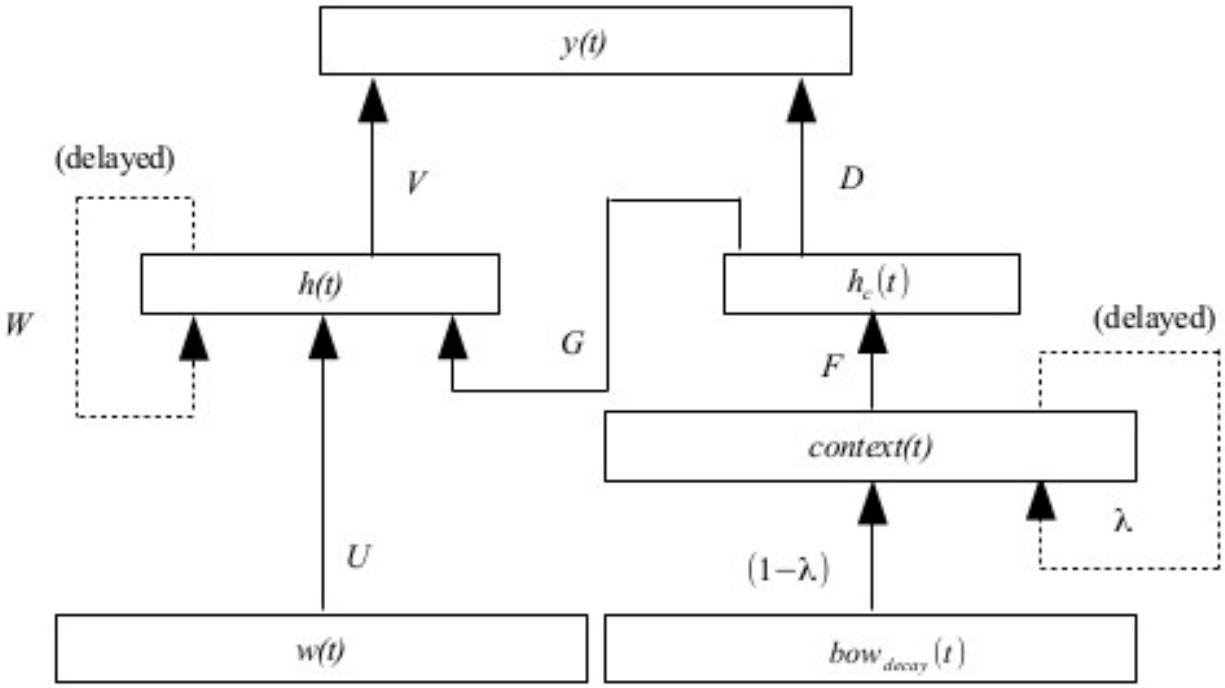
---

(1) M.Kurimo, S.Enarvi, O.Tilk, M.Varjokallio, A.Mansikkaniemi, T.Alumäe. Modeling under-resourced languages for speech recognition. Language Resources and Evaluation, 2016.

(2) M.Varjokallio, M.Kurimo, S.Virpioja. Class n-gram models for very large vocabulary speech recognition of Finnish and Estonian. Proc. SLSP 2016.

(3) A.Haidar, M.Kurimo. Recurrent Neural Network Language Model With Incremental Updated Context Information Generated Using Bag-of-Words Representation. Proc. Interspeech 2016.

An example of an extended RNN LM: Here long context is used as a sliding bag of words (bow) via a small context layer. Improves models and saves parameters (WSJ task) (1).



(1) A.Haidar, M.Kurimo. Recurrent Neural Network Language Model With Incremental Updated Context Information Generated Using Bag-of-Words Representation. Proc. Interspeech 2016.

(2) S.Enarvi, M.Kurimo. TheanoLM - An Extensible Toolkit for Neural Network Language Modeling. Proc. Interspeech 2016.

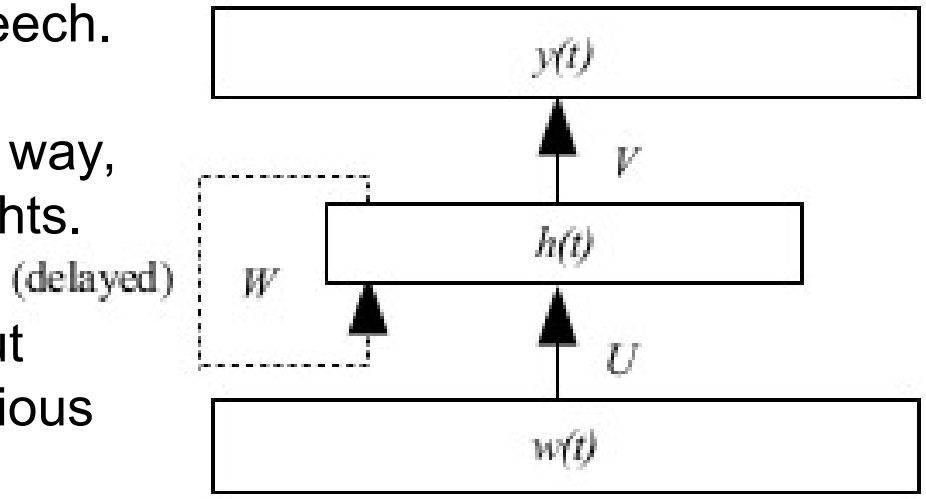
Recurrent Neural Network LM

A statistical LM that gives a probability distribution of the next word in speech.

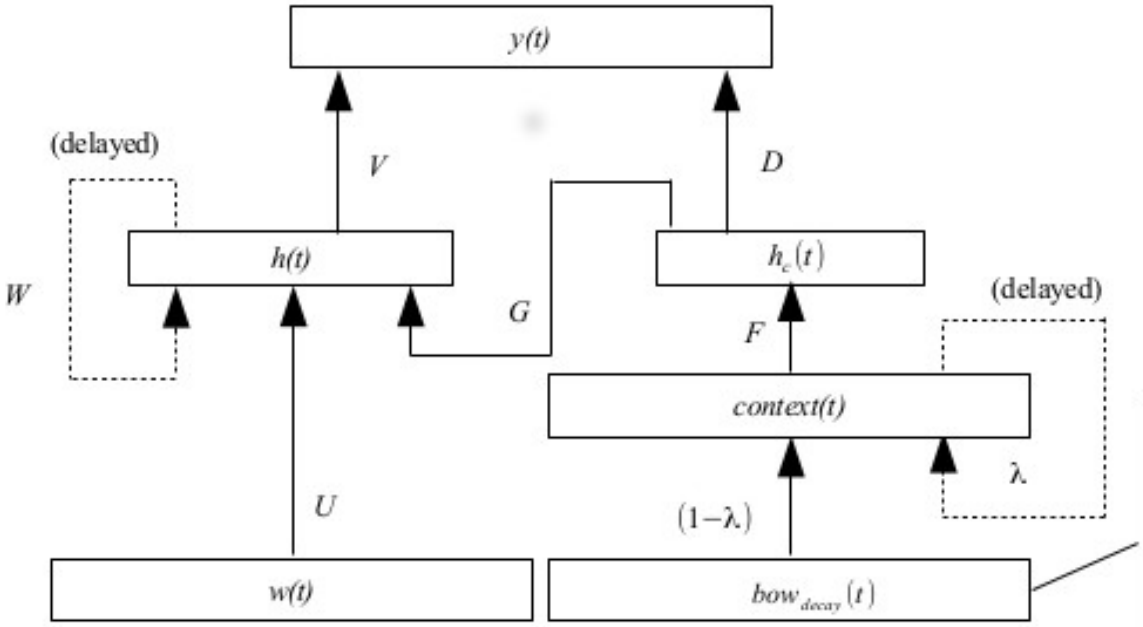
Represents words in a distributed way, as non-linear combination of weights.

Remembers history by taking input from the hidden states of the previous time steps.

Trained by stochastic gradient descent with backpropagation through time algorithm.



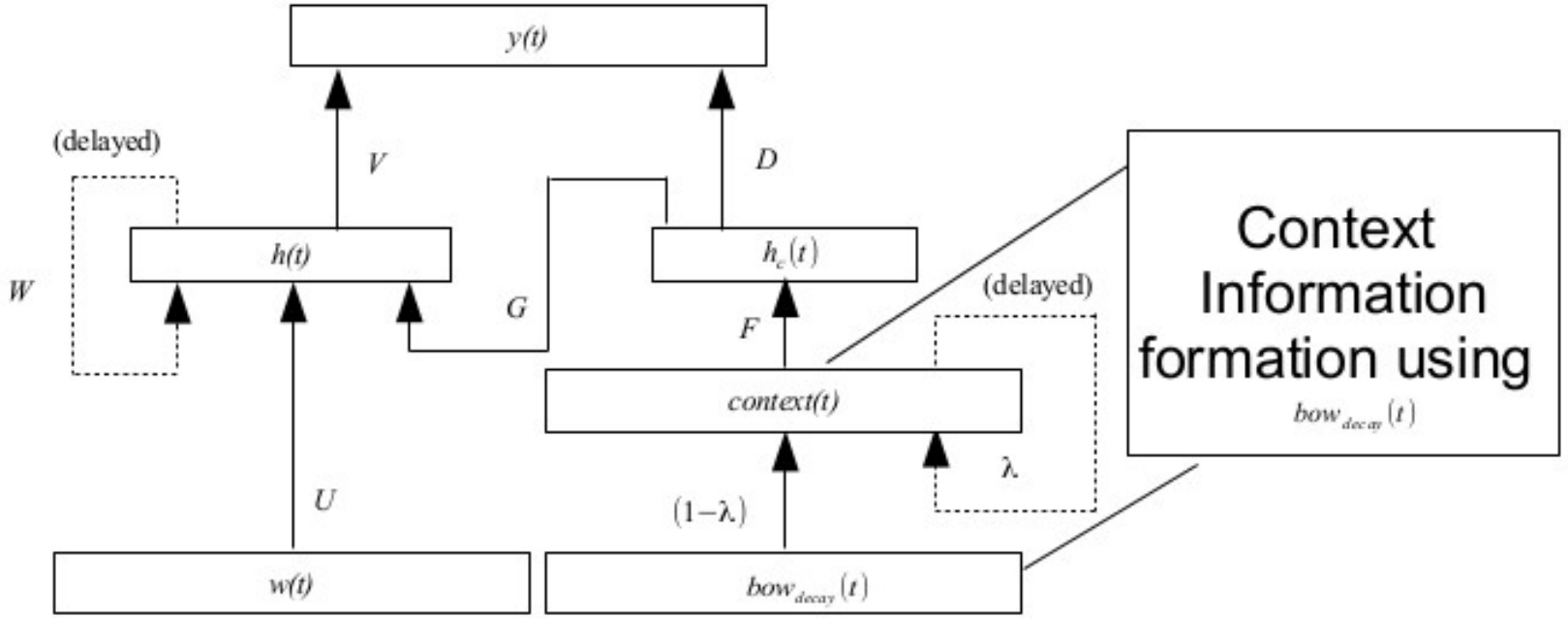
Proposed RNN-BOW LM: new input



BOW decay Representation for sliding window of size *L* words

$$bow_{decay}(t) = \sum_{i=0}^{L-1} \gamma^i w(t-i)$$

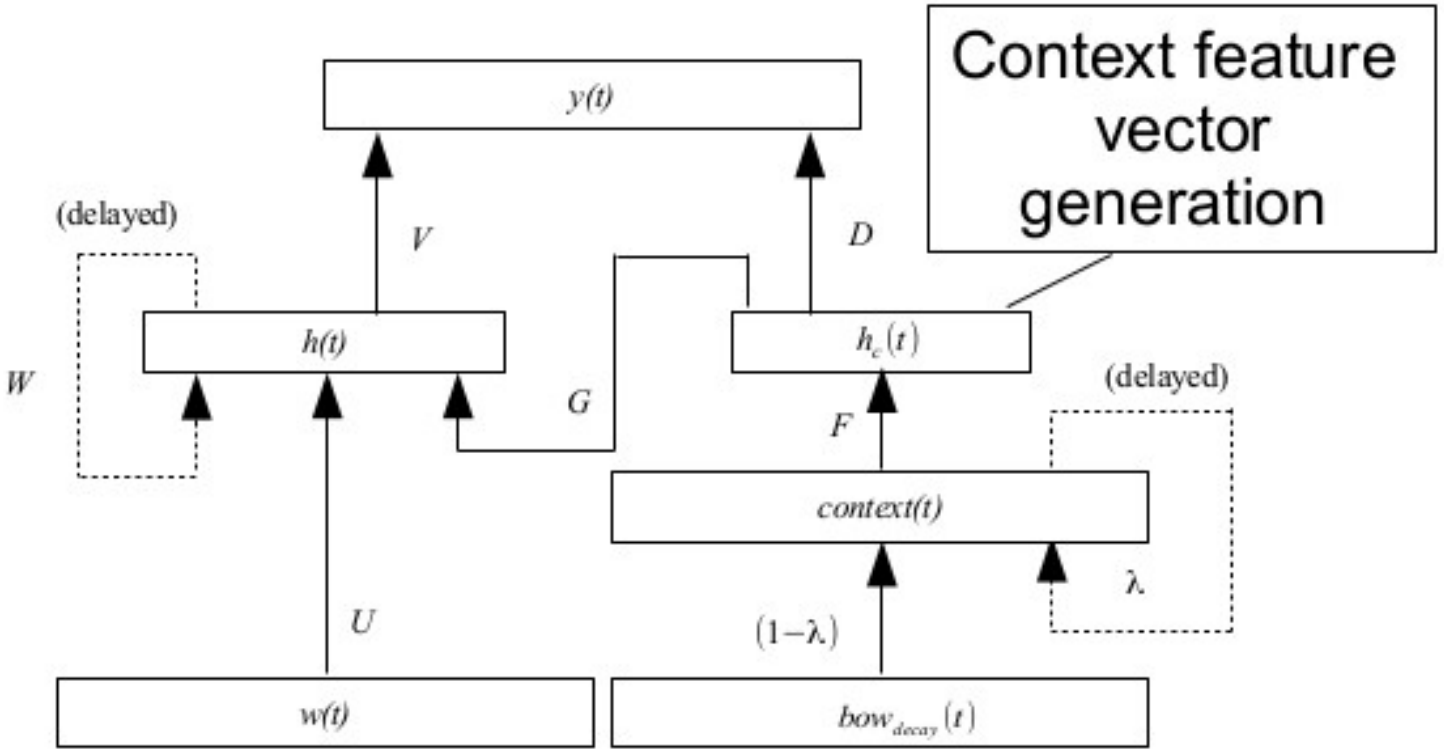
Proposed RNN-BOW LM: linear context vector



The context information vector $context(t)$ is updated as:

$$context(t) = \lambda context(t-1) + (1-\lambda) bow_{decay}(t)$$

Proposed RNN-BOW LM: non-linear context layer



$$h_c(t) = f(F context(t))$$

where $f()$ is sigmoid and F is a weight matrix

Computing the output of RNN-BOW LM

The **hidden layer** h takes input from $h(t-1)$, word $w(t)$ and the **context feature layer** $h_c(t)$:

$$h(t) = f(Uw(t) + Wh(t-1) + Gh_c(t)) \tag{8}$$

where $f()$ is sigmoid and U, W and G are weight matrices

The **output layer** y takes input from $h(t)$ and $h_c(t)$:

$$y(t) = g(Vh(t) + Dh_c(t)) \tag{9}$$

where $g()$ is soft-max and V and D are weight matrices

Language Model	$H(H_c)$	PPL	WER
KN5	-	248.0	12.8
RNN	200 (-)	226.2	12.0
RNN-BOW	190 (10)	**218.8**	**11.7**
RNN+KN5	200 (-)	191.6	11.8
RNN-BOW+KN5	190 (10)	**183.0**	**11.3**

Word error rate (WER) % and perplexity (PPL) on 1 M words Wall Street Journal speech corpus with and without class layer.

Language Model	$H(H_c)$	PPL	WER
KN5	-	248.0	12.8
RNN	200 (-)	215.6	12.0
RNN-BOW	190 (10)	**207.0**	**11.7**
RNN+KN5	200 (-)	183.4	11.7
RNN-BOW+KN5	190 (10)	**176.6**	**11.1**

RNN-BOW requires less parameters and training, but beats RNN significanly.

A.Haidar, M.Kurimo. *Recurrent Neural Network Language Model With Incremental Updated Context Information Generated Using Bag-of-Words Representation.* Proc. Interspeech 2016.

ASR demos today

1. Raw transcription of speech:
    1. Parliament sessions
    2. Television programs
2. Dictation and personal speech recordings:
    1. Offline ASR service at FIN-CLARIN and AaltoASR
       http://tinyurl.com/aaltoasr
    2. Online ASR demo
3. Speech-to-speech machine translation:
    1. Travel phrases (EMIME demo)

\+ Audio Description by Automatic Multimodal Content Analysis (ADAMCA project)

Contact:

Mikko Kurimo *mikko.kurimo@aalto.fi*

http://spa.aalto.fi/en/research/research_groups/speech_recognition/demos/

http://tinyurl.com/aaltoasr

Demos in YouTube:

https://www.youtube.com/channel/UCY4NOvOgKz9-x7rR_kkb51Q