Google Speech Processing from Mobile to Farfield

Michiel Bacchiani

Tara Sainath, Ron Weiss, Kevin Wilson, Bo Li, Arun Narayanan, Ehsan Variani, Izhak Shafran, Kean Chin, Ananya Misra, Chanwoo Kim,
…and many others in the speech and related teams

Google Inc.

2006

2011

TECHNOLOGY | INNOVATION, THE INTERNET, GADGETS, AND MORE. | APRIL 6 2011 4:36 PM

Now You're Talking!

Google has developed speech-recognition technology that actually works.

By Farhad Manjoo

Google Speech Group
Early Days "Mobile"

- Speech group started in earnest in 2005
- Build up our own technology, first application launched in April 2007
- Simple directory assistance
- Early view of what a "dialer" could be

Google Speech Group
Early Days Voicemail

Launched early 2009 as part of Google Voice

Voicemail transcription:

- navigation
- search
- information extraction

Google Speech Group Early Days YouTube

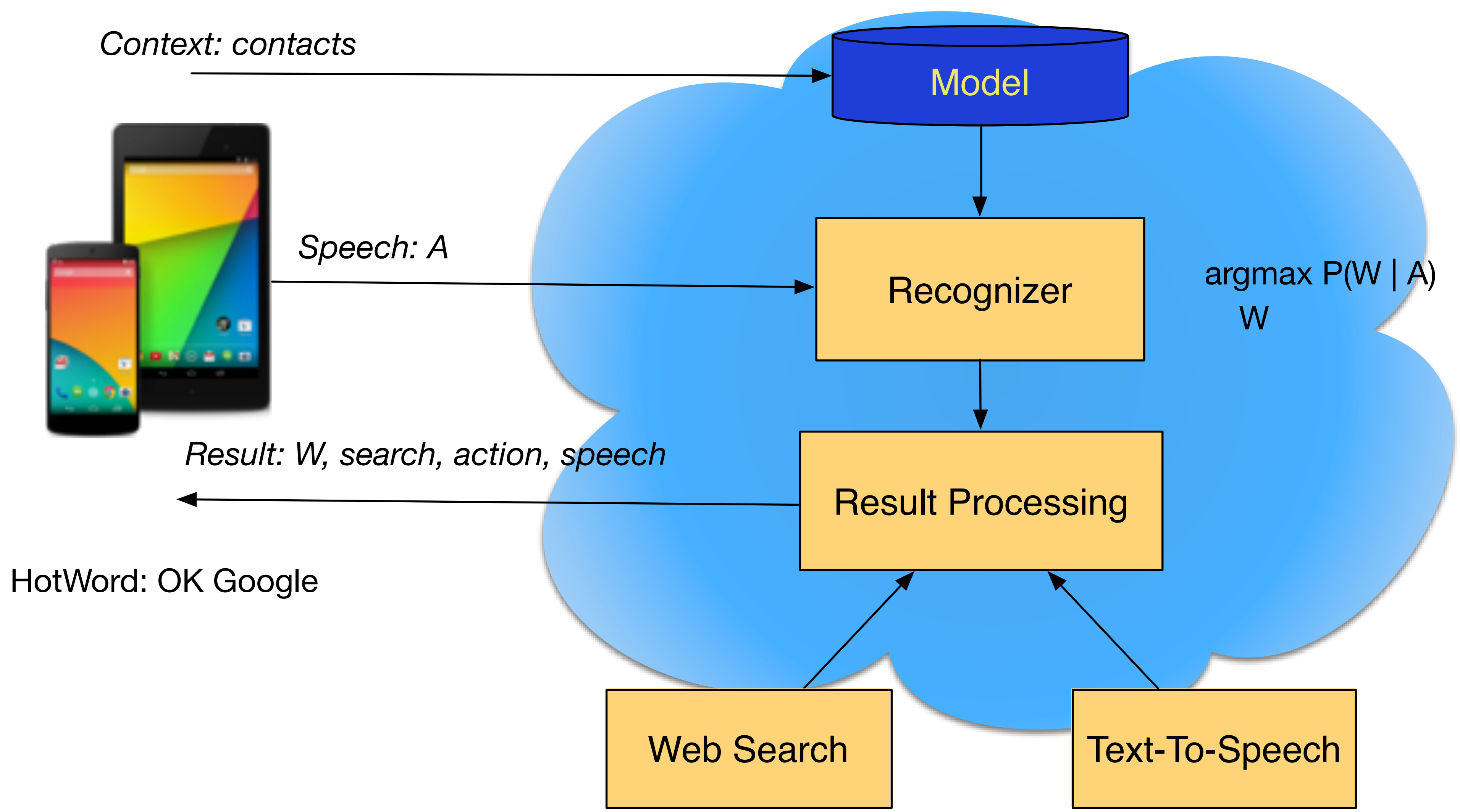
Launched early 2010

- automatic captioning
- translation
- editing, “time sync”
- navigation

The Revolution

- Early speech applications had some traction but nothing like the engagement we see today
- The 2007 launch of smartphones (iPhone and Android) was a revolution and dramatically changed the status of speech processing
- Our current suite of mobile applications is launched in 60+ languages and processes about a century of speech each day

Mobile Application Overview



Recognition Models

Multi-lingual

Language Model	Domain/Text Norm: 7:15AM $3.22	P(W)	Lexical
	Dynamic Language Model Biasing Dynamic Lexical Items: Contact Names		↓
Lexicon	Size/Generalization: goredforwomen.org		
Acoustic Model	Acoustic Units/Context/Distribution Estimation	P(A | W)	Acoustic

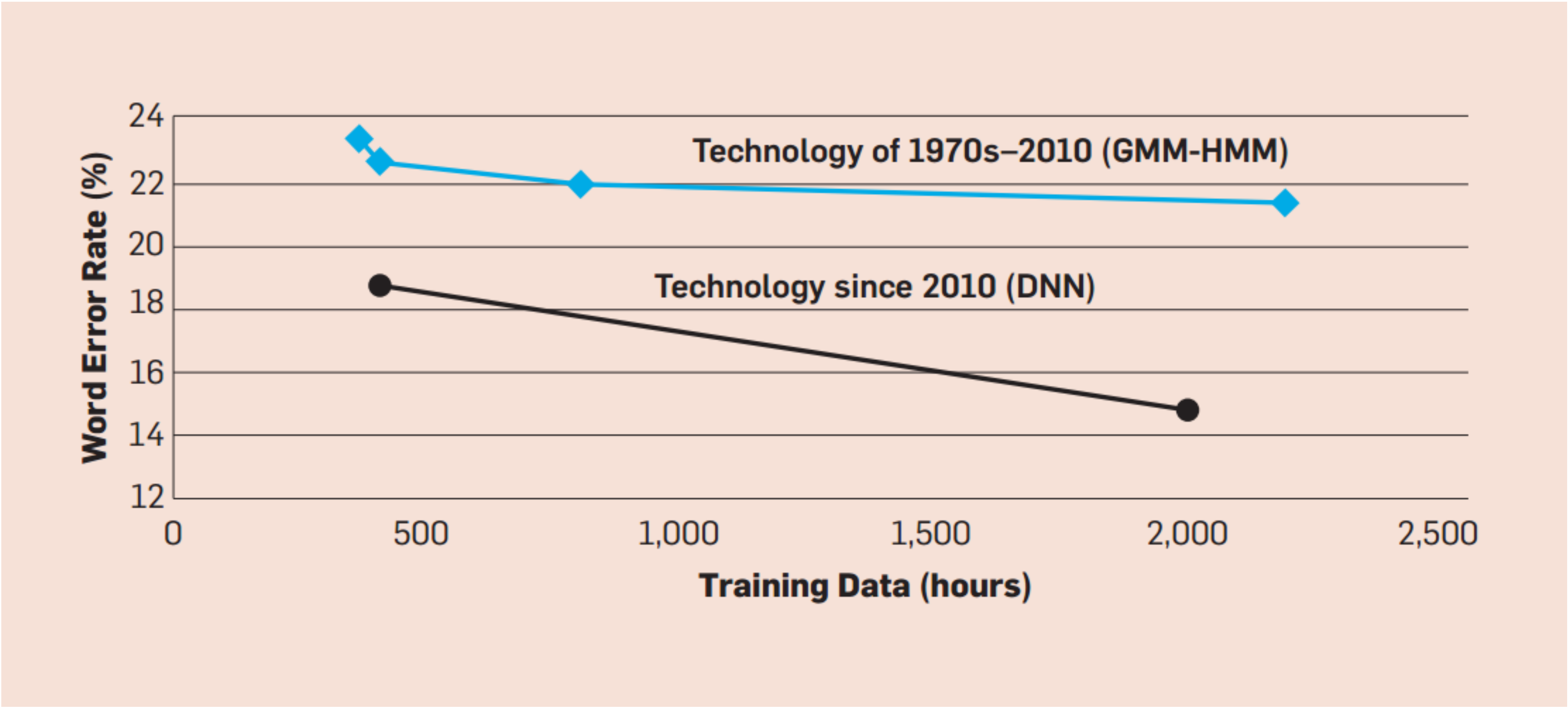
Deep Neural Networks

Finite State Transducers

App Context vs. Technology

Mobile makes use of accurate speech recognition compelling

Large volume use improves statistical models



Xuedong Huang, James Baker and Raj Reddy, *"A Historial Perspective of Speech Recognition,"* Communications of the ACM, January 2014, Vol. 57, No 1.

DNN Technical Revolution

First resurgence

- Abdel-rahman Mohamed, George Dahl and Geoffrey Hinton *"Deep belief networks for phone recognition,"* In NIPS Workshop on Deep Learning for Speech Recognition and Related Applications. 2009
- Abdel-rahman Mohamed and Geoffrey Hinton *"Phone recognition using Restricted Boltzmann Machines,"* In the proceeding of ICASSP 2010

Large Vocabulary

- Dahl, Mohamed and Jaintly intern at Microsoft, IBM and Google and show LVCSR applicability

First Industry LVCSR Results

- Microsoft shows gains on the SwitchBoard task.
    - Frank Seide, Gang Li, and Dong Yu, "*Conversational Speech Transcription Using Context-Dependent Deep Neural Networks*," In the proceedings of Interspeech 2011.

Google uses DNN in its products

2009

2010

2011

2012

DNN vs. GMM

	Model Type	WER (%)	Training Size (hours)	GPU Training Time (hours/epoch)	Hidden Layers	Number of States
VoiceSearch	GMM	16.0	5780	321	4x2560	7969
	DNN	12.2				
YouTube	GMM	52.3	1400	55	4x2560	17552
	DNN	46.2				

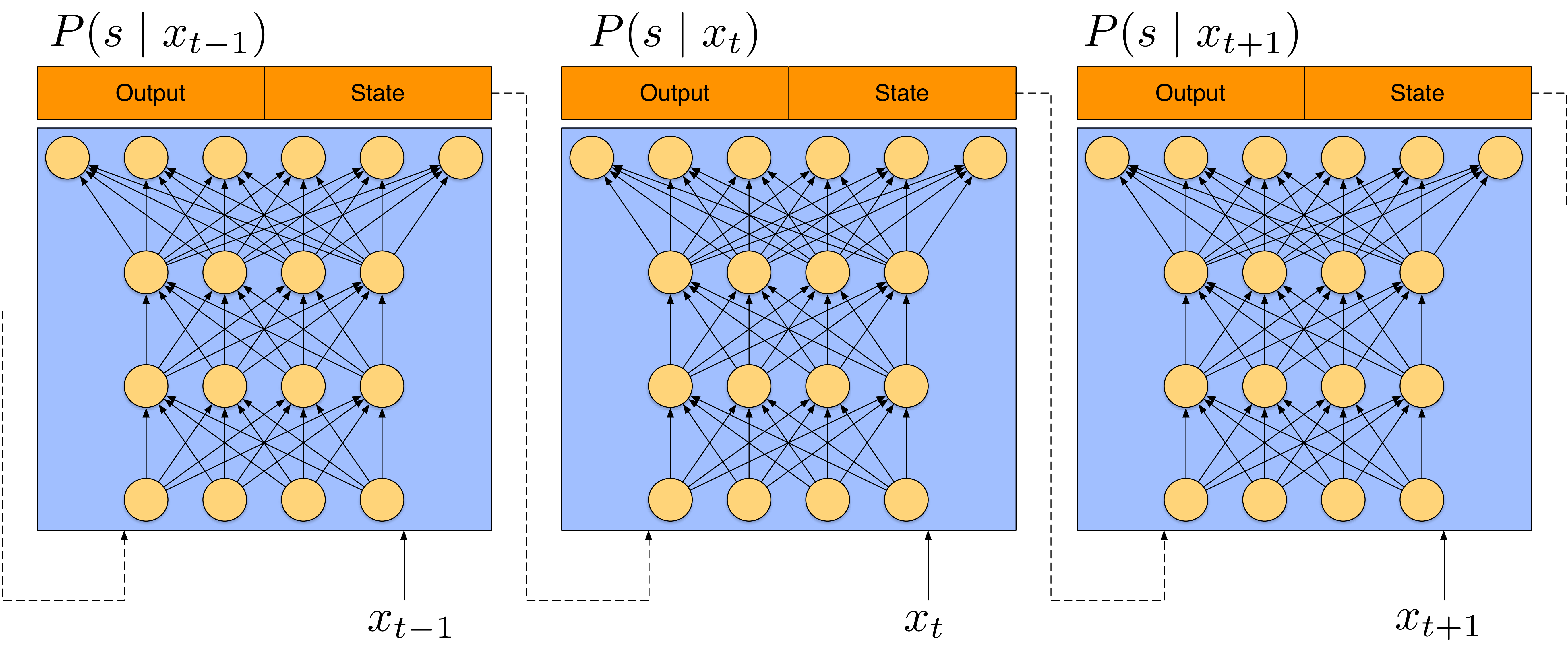
DistBelief CPU training allows speed ups of 70 times over a single CPU and 5 times over a GPU.

Train a 85M parameter system on 2000 hours, 10 epochs in about 10 days.

Jeffrey Dean, Greg S. Corrado, Rajat Monga, Kai Chen, Matthieu Devin, Quoc V. Le, Mark Z. Mao, Marc'Aurelio Ranzato, Andrew Senior, Paul Tucker, Ke Yang, Andrew Y. Ng, "Large Scale Distributed Deep Networks," in the proceeding of NIPS (2012)

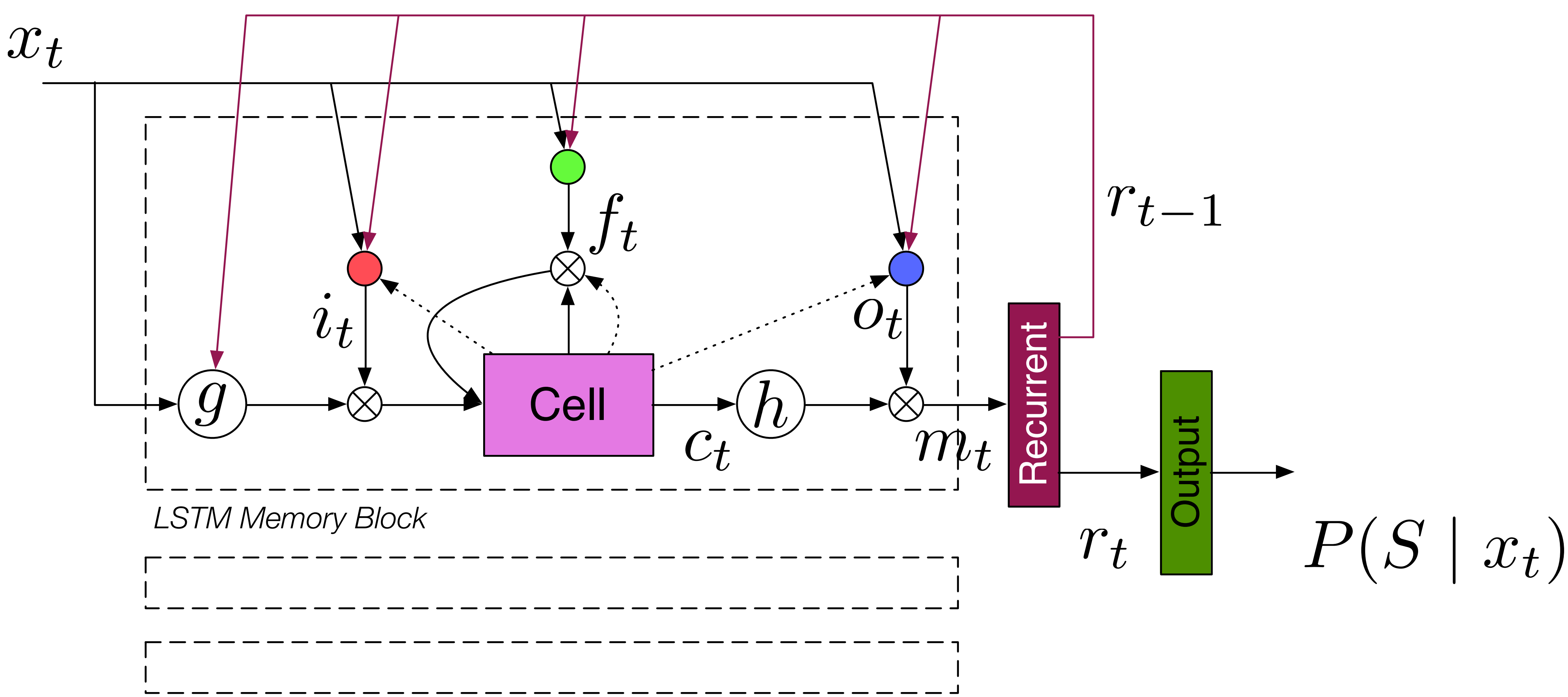
Using a Sequence Model

The DNN can be trained with a sequence objective but it still bases it estimation on the current observation alone



Long Short Term Memory

With a moderate increase in complexity, get much better behavior of BPTT training.



Training LSTMs with CE

8x2560 hidden layer DNN reaches 11.3% WER with CE training, 10.4% with sequence training

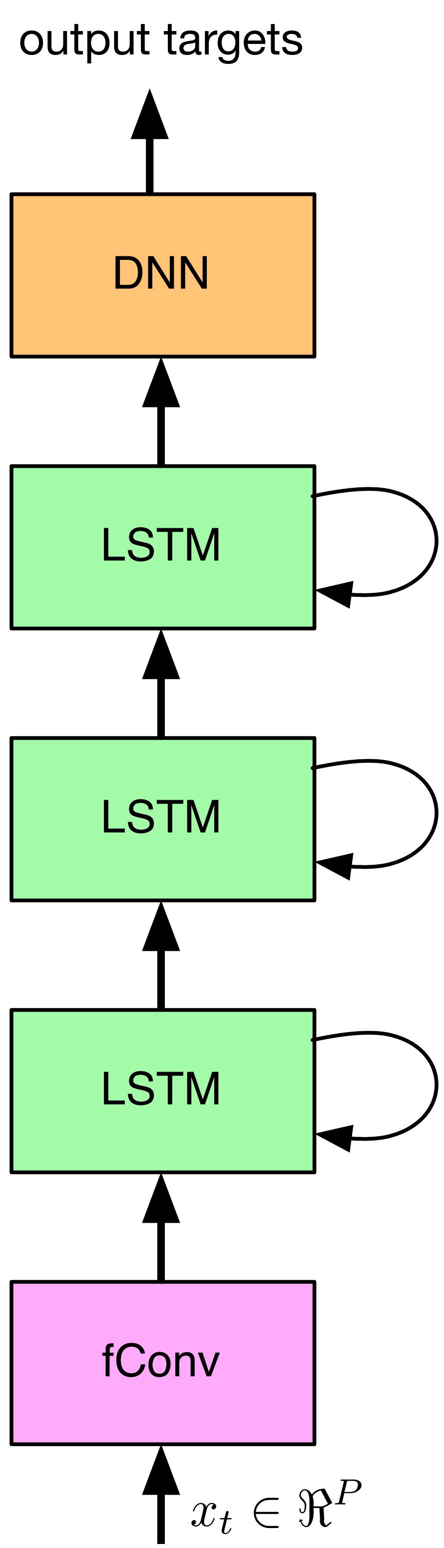
Cells	Projection	Depth	Parameters	WER(%)
750		1	13M	12.4
385		7	13M	11.2
600		2	13M	11.3
440		5	13M	10.8
840		5	37M	10.9
2048	512	1	13M	11.3
800	512	2	13M	10.7
1024	512	3	20M	10.7
2048	512	2	22M	10.8
6000	800	1	36M	11.8

Sequence Training LSTMs

- Since the LSTM model has a state to model the sequence, it will "learn the language model" if trained with a CE criterion.
- Sequence training will focus its learning on the acoustic sequence model.

Model Type	DNN		LSTM	
Objective	CE	Sequence	CE	Sequence
WER	11.3	10.4	10.7	9.8

CLDNNs



- Added accuracy improvements from combining layers of different types.

2000 hour clean training set,
20 hour clean test set

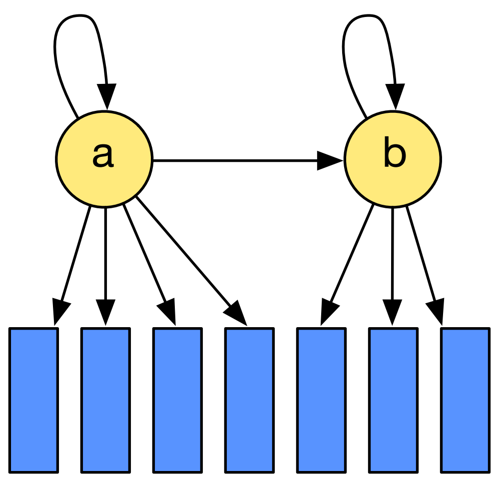
	CE	Sequence
LSTM	14.6	13.7
CLDNN	13.0	13.1

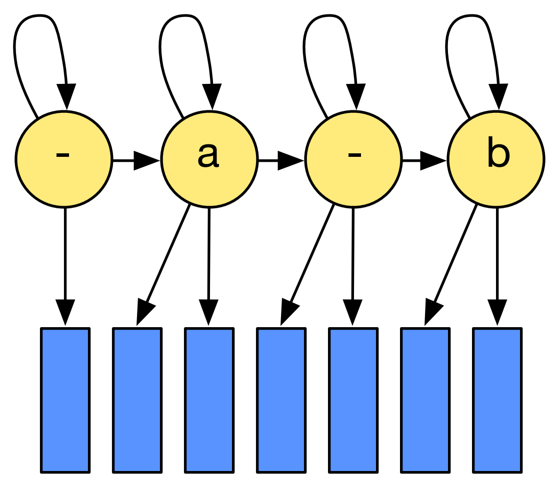
2000 hour MTR training set,
20 hour noisy test set

	CE	Sequence
LSTM	20.3	18.8
CLDNN	19.4	17.4

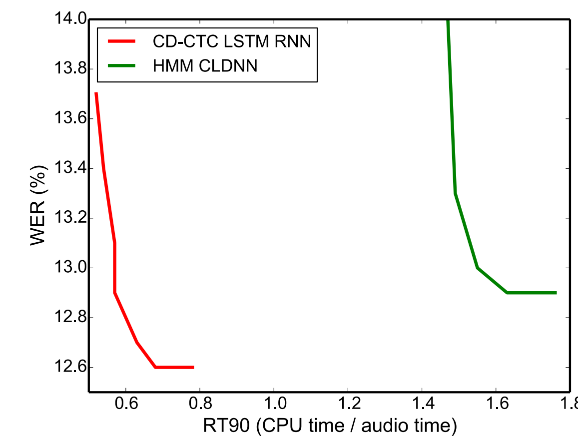
CTC and Low Frame Rate

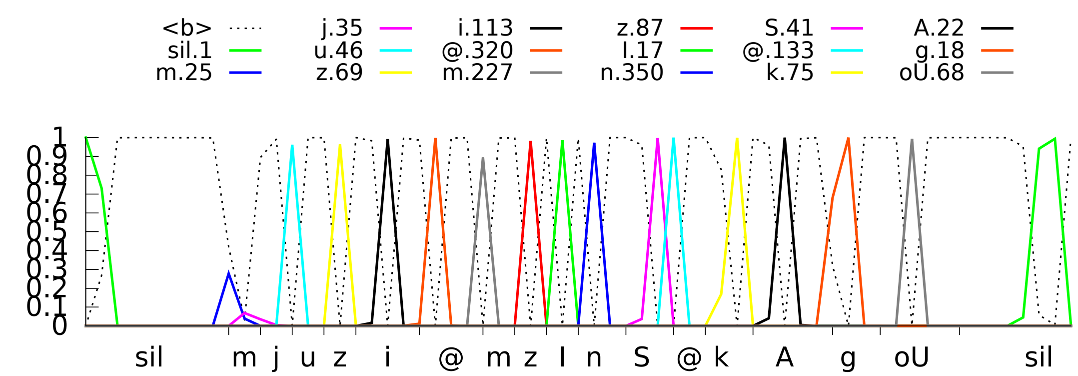
$$P(X) = \sum_{s \in \{S\}} P(X, s)$$

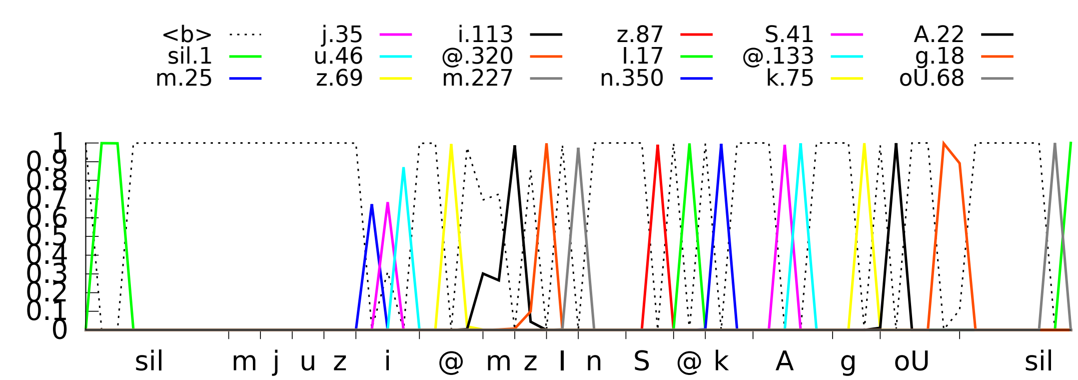




$$\sum_{s \in \{S\}} \prod_{t=1}^{T} P(x_t|s_t) P(s_t|s_{t-1}) \qquad \sum_{s \in \{S\}} \prod_{t=1}^{T} P(s_t|X)$$

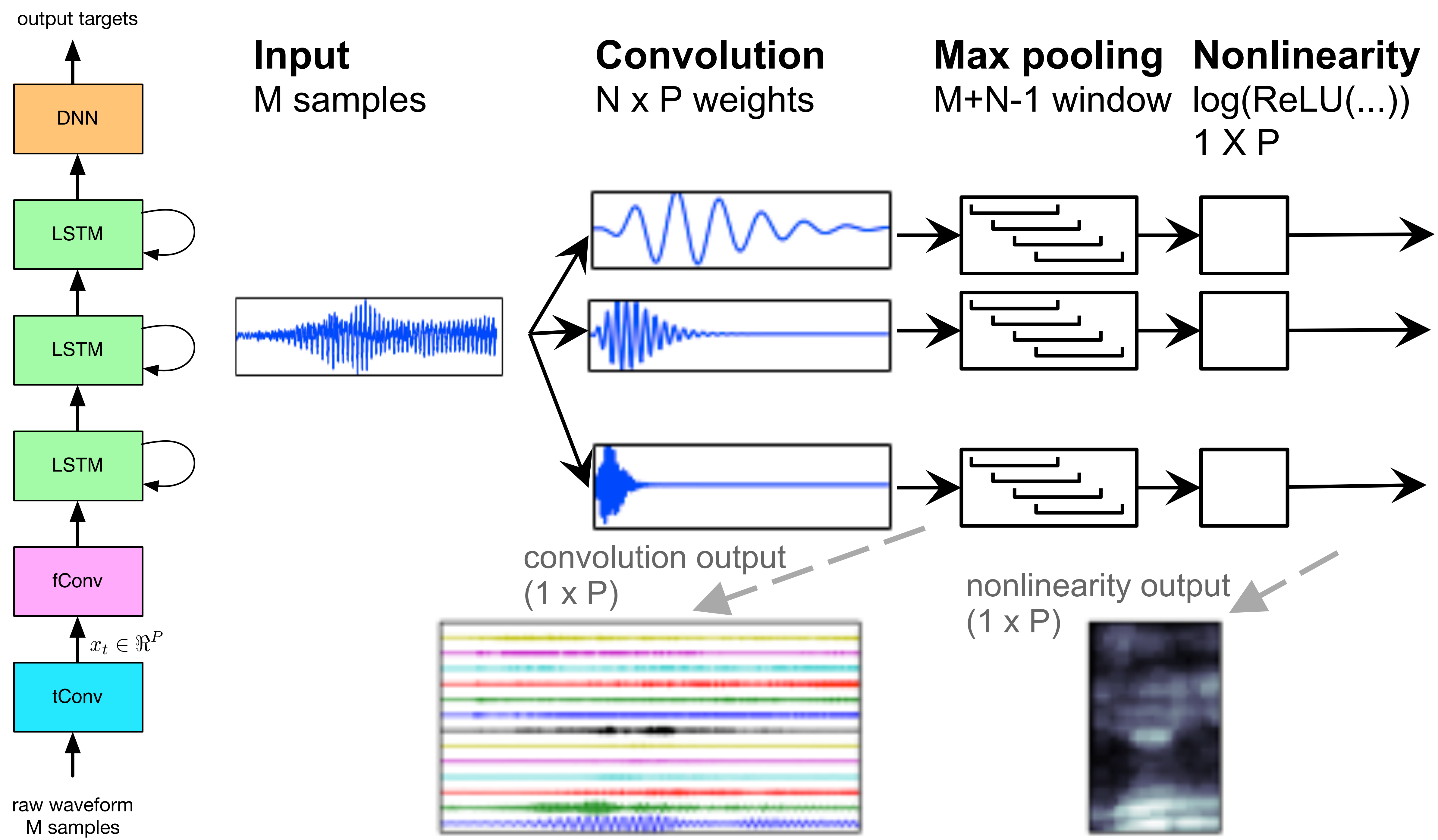




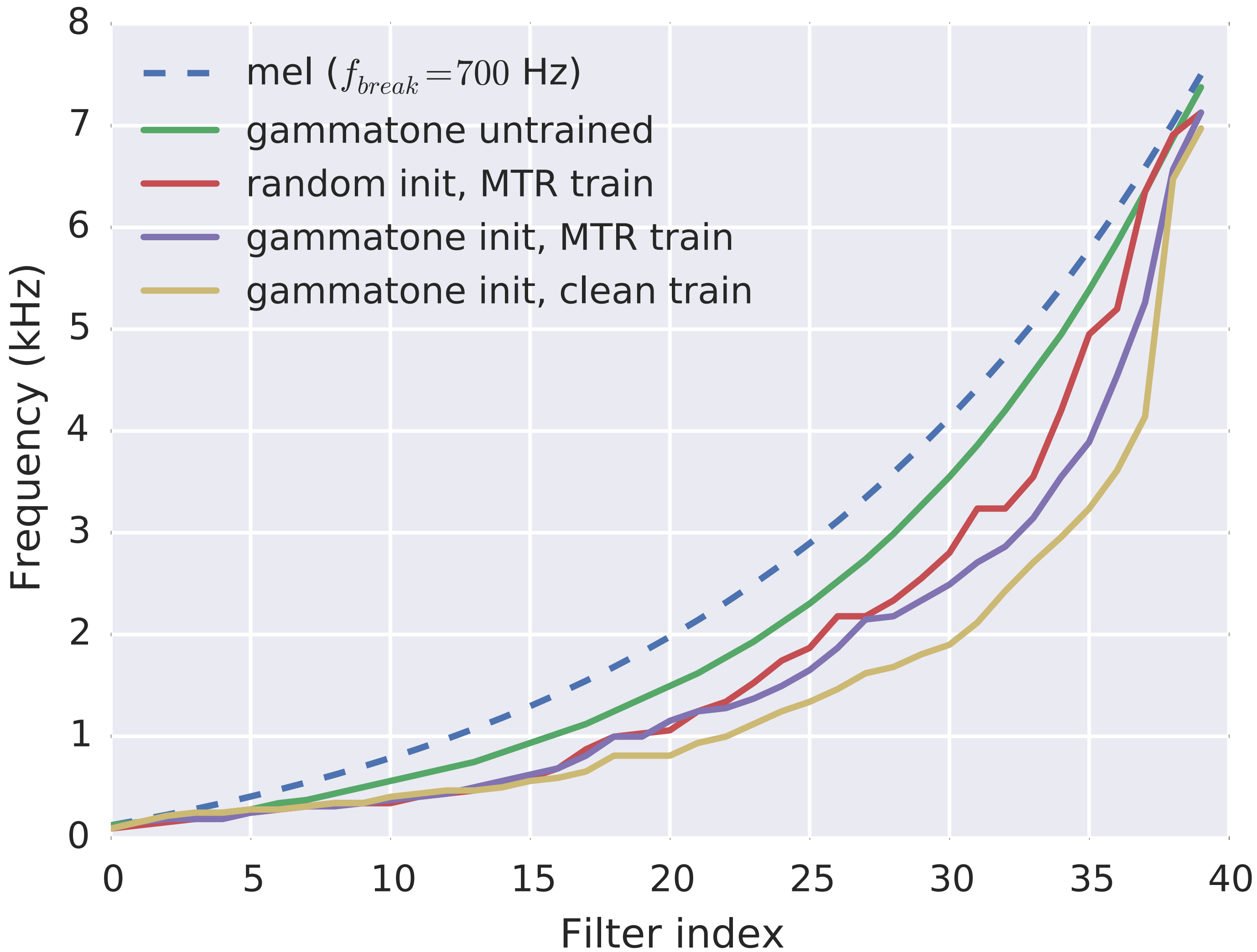


100 ms alignment constraint

Raw Waveform Models



Raw Waveform Performance



Feature	Model	WER
Log-mel	C1L3D1	16.2
Raw	C1L3D1	16.2
Log-mel	L3D1	16.5
Raw	L3D1	16.5
Raw	L3D1 rnd	16.5
Log-mel	D6	22.3
Raw	D6	23.2

Farfield

- A new way for people to interact with the internet
- More natural interface in the home
- More social

- User expectations based on phone experience
- Technically a non-trivial problem: reverb, noise, level differences

Data Approach

- New application, no prior data that is
  - Multi-channel
  - Reverberant
  - Noisy
- Lots of data from phone launched applications (maybe noisy/reverberant, but no control)
- Bootstrap approach to build a room simulator (IMAGE method) to generate “room data” from “clean data”

Training Data

- 2000 hour set from our anonymized voice search data set
- Room dimensions sampled from 100 possible configurations
- T60 reverberation ranging from 400 to 900 ms. (600ms. ave)
- Simulate an 8-channel uniform linear mic array with 2cm mic spacing
- Vary source/target speaker locations, distances from 1 to 4 meters
- Noise corruption with “daily life” and YouTube music/noise data sets
- SNR distribution ranging from 0 to 20 dB SNR

Test Data

- Evaluate on a 30k voice search utterance set, about 20 hours
- One version simulated like the training set
- Another by **re-recording**
  - In a physical room, playback the test set from a mouth simulator
  - Record from an actual mic array
  - Record speech and noise from various (different) angles
  - Post mix to get SNR variations
- The baseline is MTR trained: early work with the room simulator (DNN models) showed

  16.2% clean-clean -> 29.4% clean-noisy -> 19.6% MTR-noisy

Multi-channel ASR

- Common approach separates enhancement and recognition

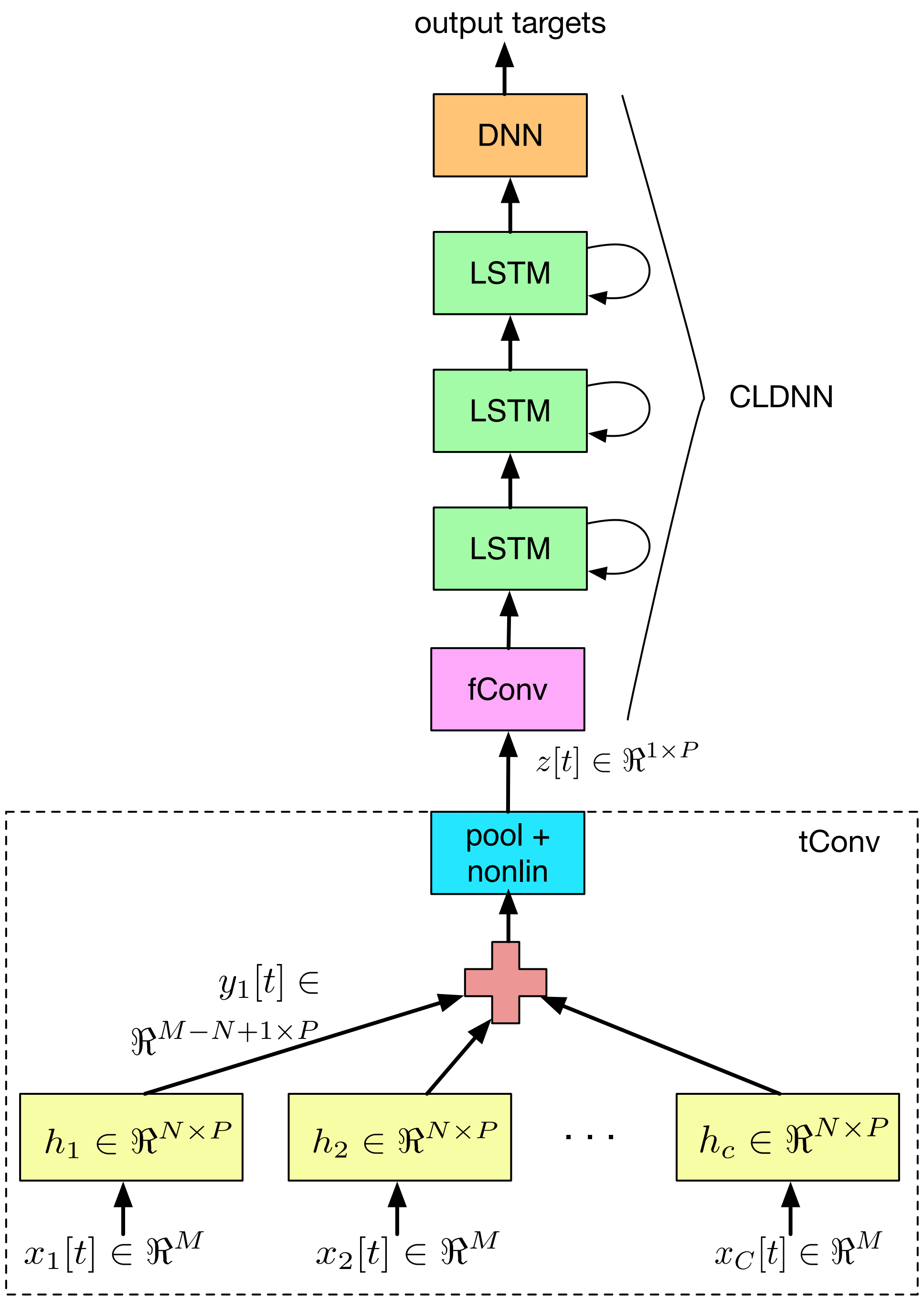
- Enhancement commonly done in localization, beamforming and postfiltering stages

- Filter-and-sum beamforming takes a steering delay from localization for the c-th channel τ_c

$$y[t] = \sum_{c=0}^{C-1} \sum_{n=0}^{N-1} h_c[n] x_c[t - n - \tau_c]$$

- Estimation is commonly based on Minimum Variance Distortionless Response (MVDR) or Multi-channel Wiener Filtering (MWF)

Raw Multi-Channel

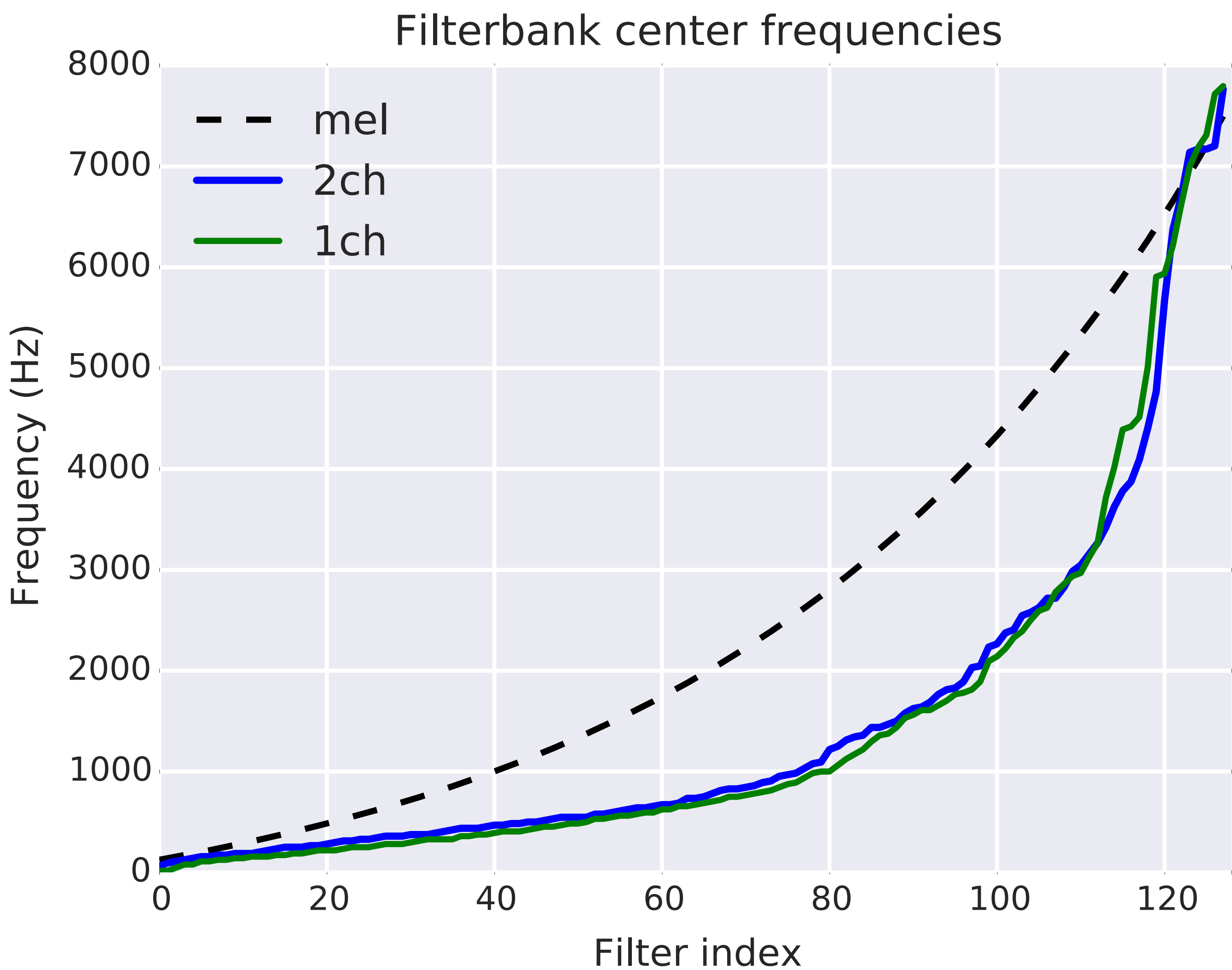


$$y^p[t] = \sum_{c=0}^{C-1} \sum_{n=0}^{N-1} h_c^p[n] x_c[t-n]$$

- Implicitly model steering delay in a bank for P multi-channel filters

- Optimize the filter parameters directly on ASR objective akin to raw waveform single channel model.

Learned Filters



Filters	2ch (14cm)	4ch (4-6-4cm)	8ch (2cm)
128	21.8	21.3	21.1
256	21.7	20.8	20.6
512	-	20.8	20.6

Removing Phase

Train a baseline system with Log-mel features and feed these as feature maps into the CLDNN

Log-mel

Filters	2ch (14cm)	4ch (4-6-4cm)	8ch (2cm)
128	22.0	21.7	22.0
256	21.8	21.6	21.7

Raw-waveform

Filters	2ch (14cm)	4ch (4-6-4cm)	8ch (2cm)
128	21.8	21.3	21.1
256	21.7	20.8	20.6

Localization

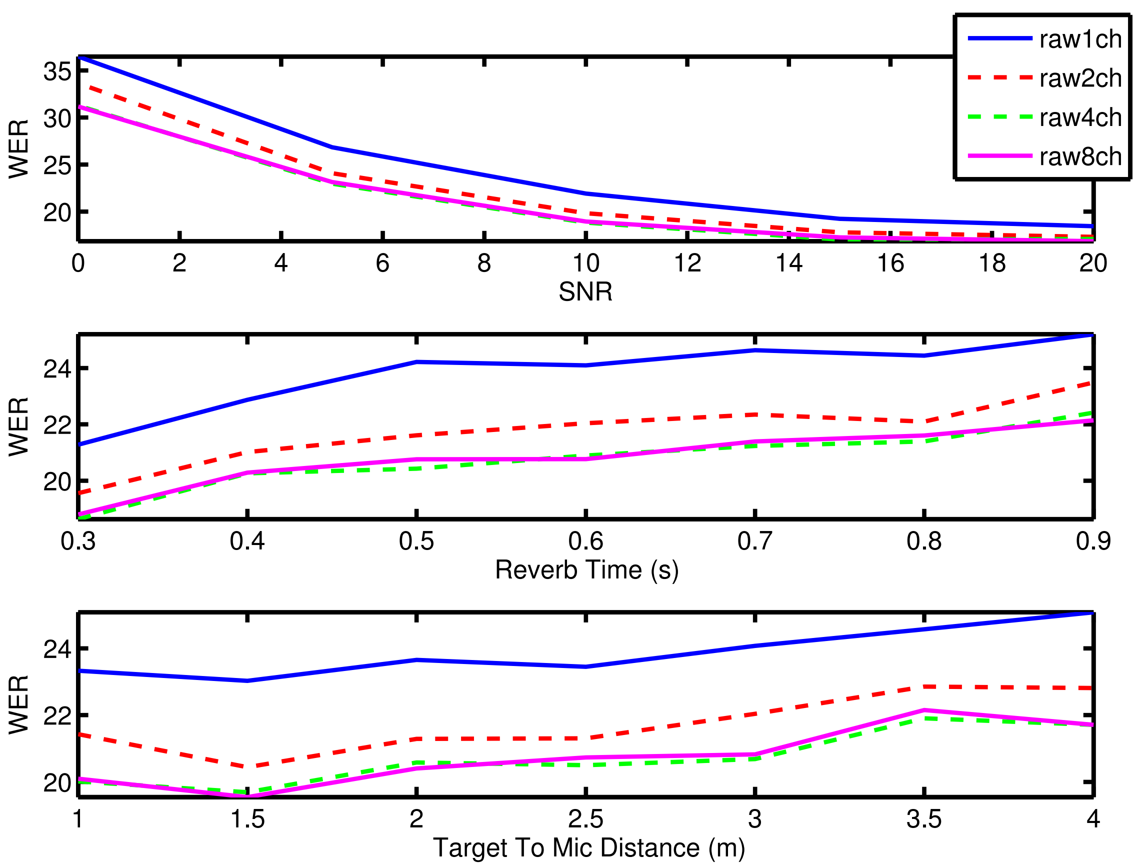
- The multi-channel raw waveform model does both beam forming as well as localization.

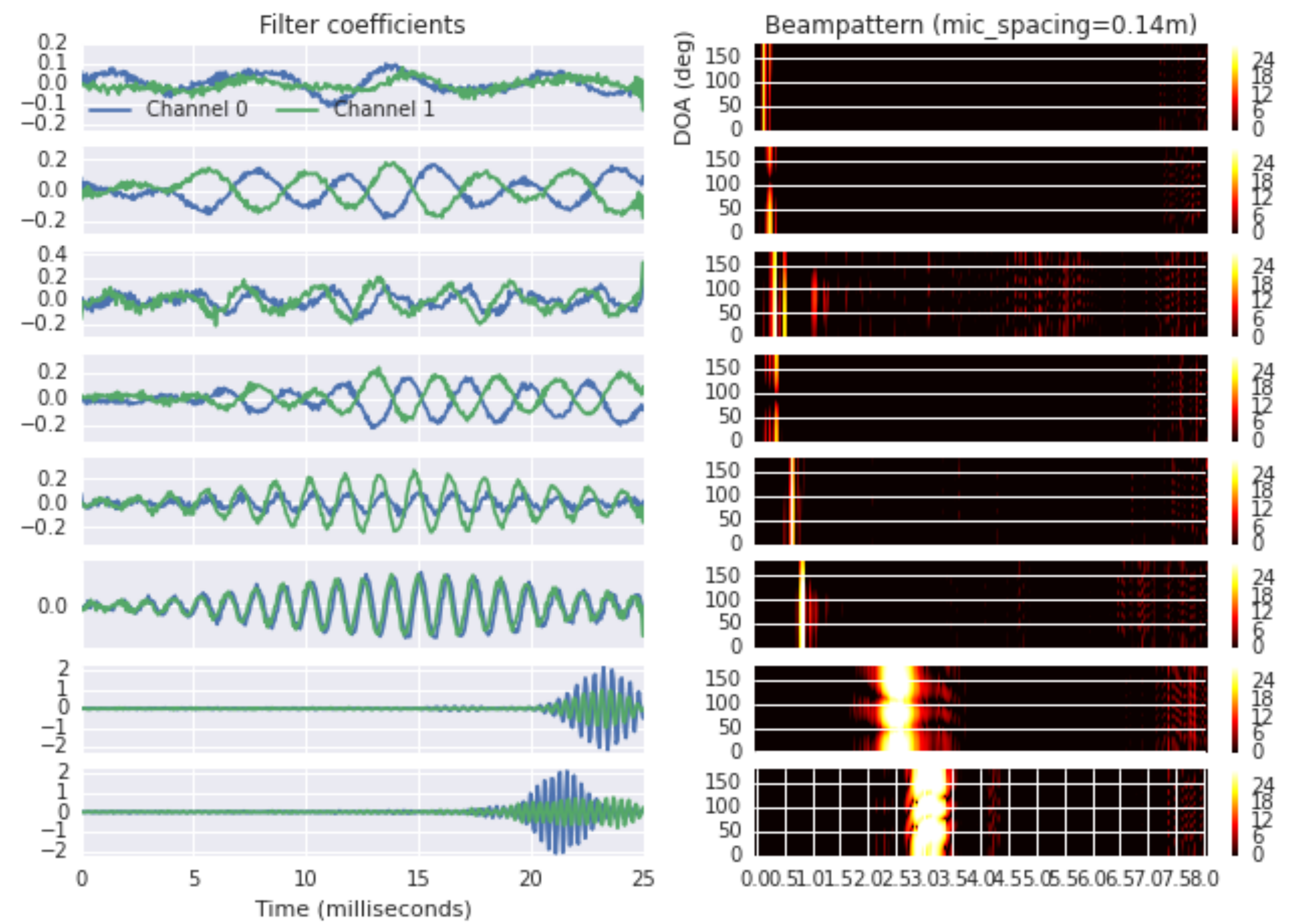
- Train a Delay-and-Sum (D+S) single channel signals with the oracle Time Delay of Arrival (TDOA)

- Train a Time Aligned Multi-channel (TAM) system where we oracle TDOA align the channel inputs.

Filters	1ch	2ch (14cm)	4ch (4-6-4cm)	8ch (2cm)
Oracle D+S	23.5	22.8	22.5	22.4
Oracle TAM	23.5	21.7	21.3	21.3
Raw, no tdoa	23.5	21.8	21.3	21.1

WER and Filter Analysis





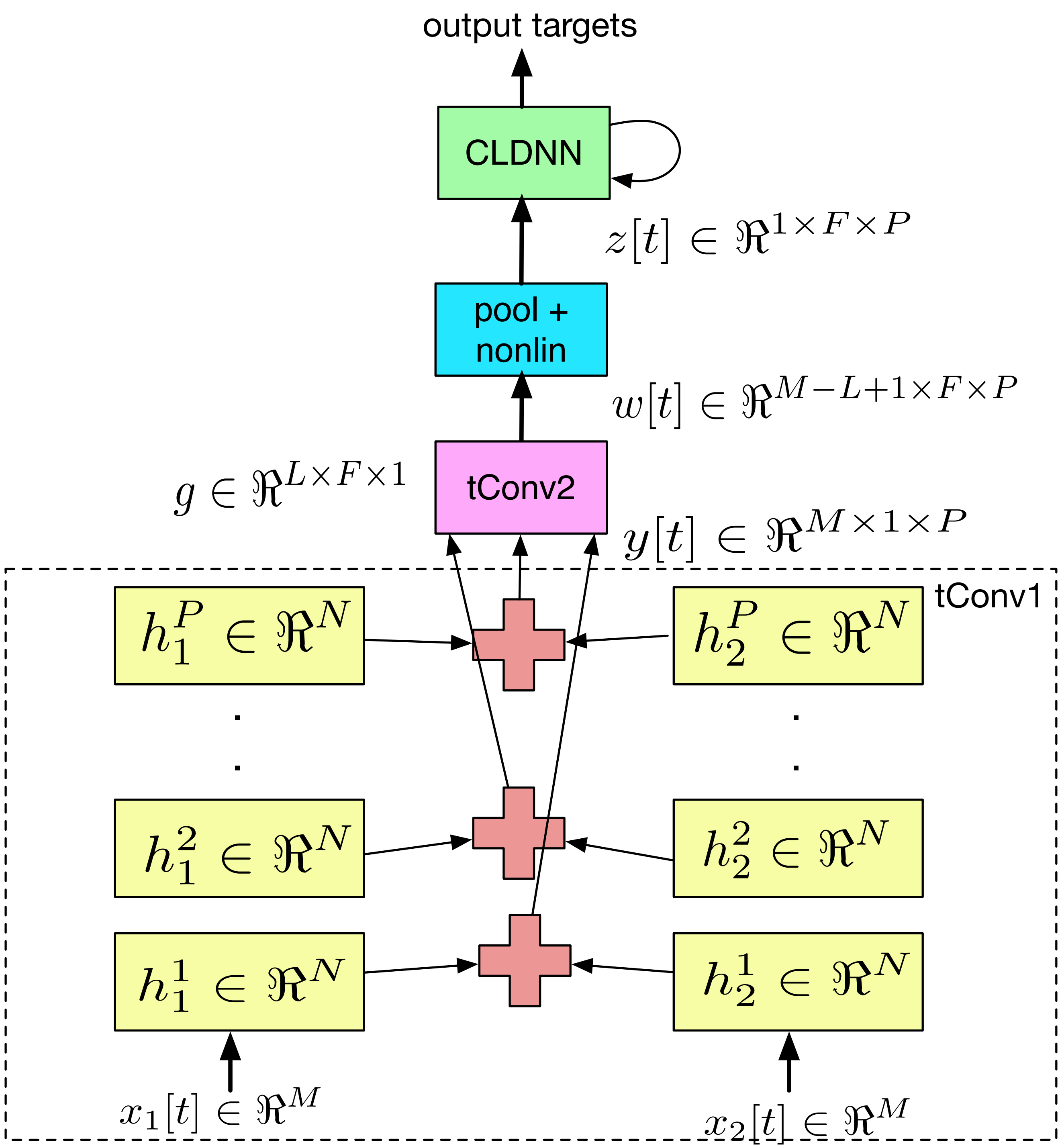
Multi-Channel Raw Waveform Summary

- Performance improvements remain after sequence training
- The raw waveform models without any oracle information do better than an MVDR model that was trained with oracle TDOA and noise

Model	WER-CE	WER-Seq
Raw 1ch	23.5	19.3
D+S, 8ch, oracle	22.4	18.8
MVDR, 8ch, oracle	22.5	18.7
raw, 2ch	21.8	18.2
raw, 4ch	20.8	17.2
raw, 8ch	20.6	17.2

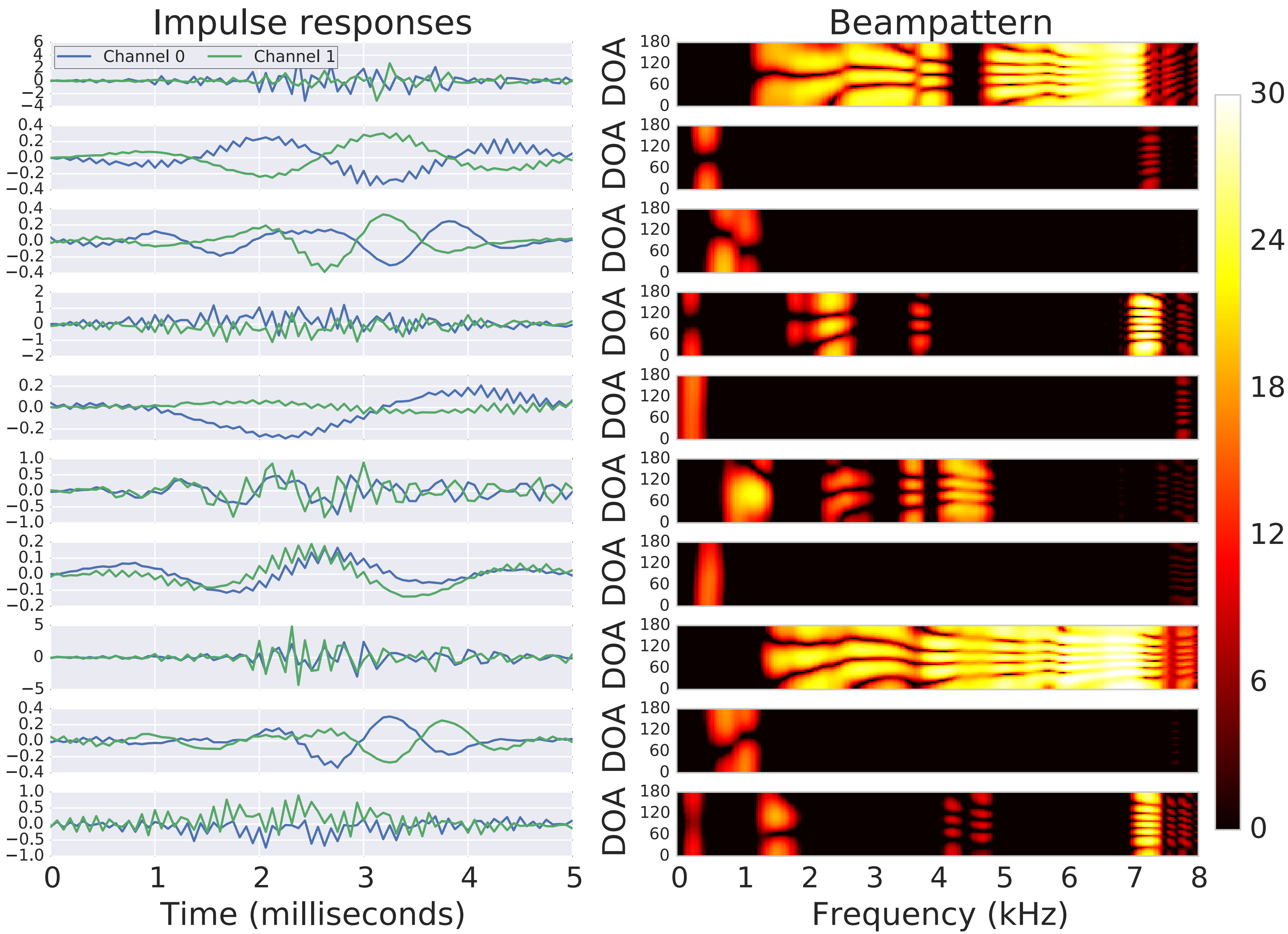
All systems 128 filters

Factored Multi-Channel Raw Waveform



- In a first convolutional layer, apply filtering for P look-directions.
- Small number of taps to encourage learning of spatial filtering
- In a second convolutional layer, use a larger number of taps for frequency resolution. Tie filter parameters between look directions

Learned Filters



Performance of Factored Models

- Factored performance improves on unfactored with increasing number of spatial filters

- Fixing the spatial filters to be D+S shows inferior

# Spatial Filters	WER
2ch, unfactored	21.8
1	23.6
3	21.6
5	20.7
10	20.8

tConv1	WER
fixed	21.9
trained	20.9

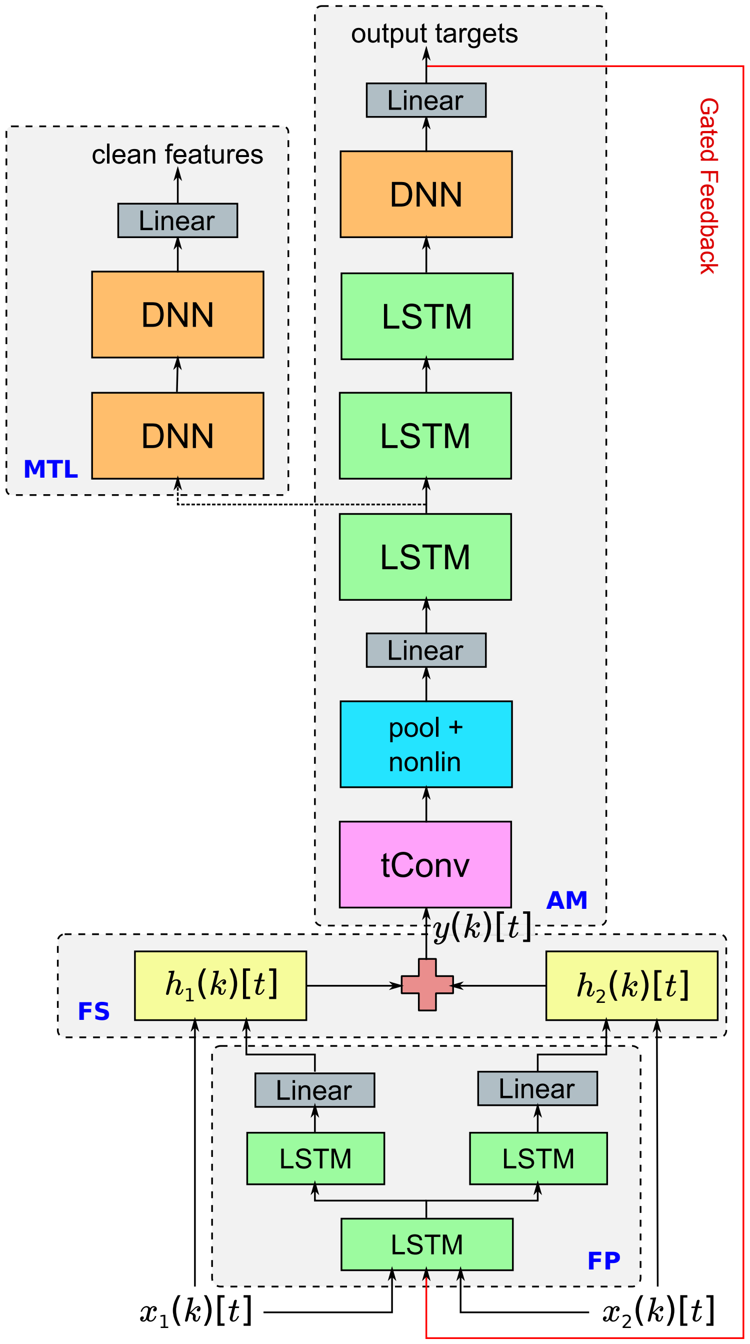
P=5 “look directions”

Multi-Channel Factored Raw Waveform Summary

- Performance improvements remain after sequence training

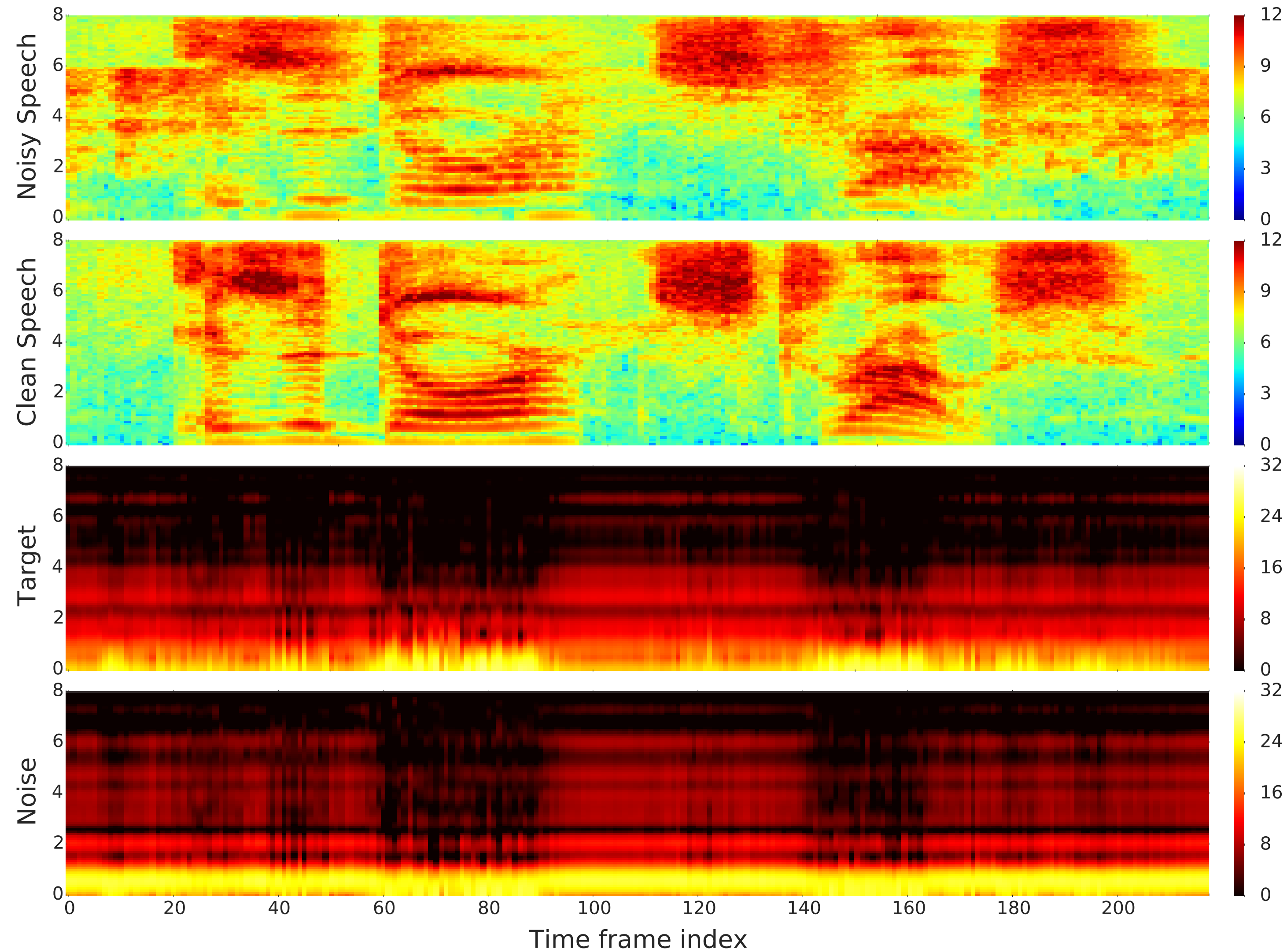
Model	WER-CE	WER-Seq
unfactored, 2ch	21.8	18.2
factored, 2ch	20.4	17.2
unfactored 4ch	20.8	17.2
factored 4ch	19.6	16.3

Neural network Adaptive Beamforming (NAB)



- An alternative to relying on factoring is to make the beamforming an adaptive process.
- Use an LSTM with the channel inputs as well as a previous prediction feedback signal to predict the filter-and-sum parameters of the incoming signals.
- Found additional gains from applying Multi-Target Learning.

NAB Results



Model	WER-CE	WER-Seq	Params(M)	MultAdd(M)
factored	20.4	17.1	18.9	35.1
NAB	20.5	17.2	24.0	28.8

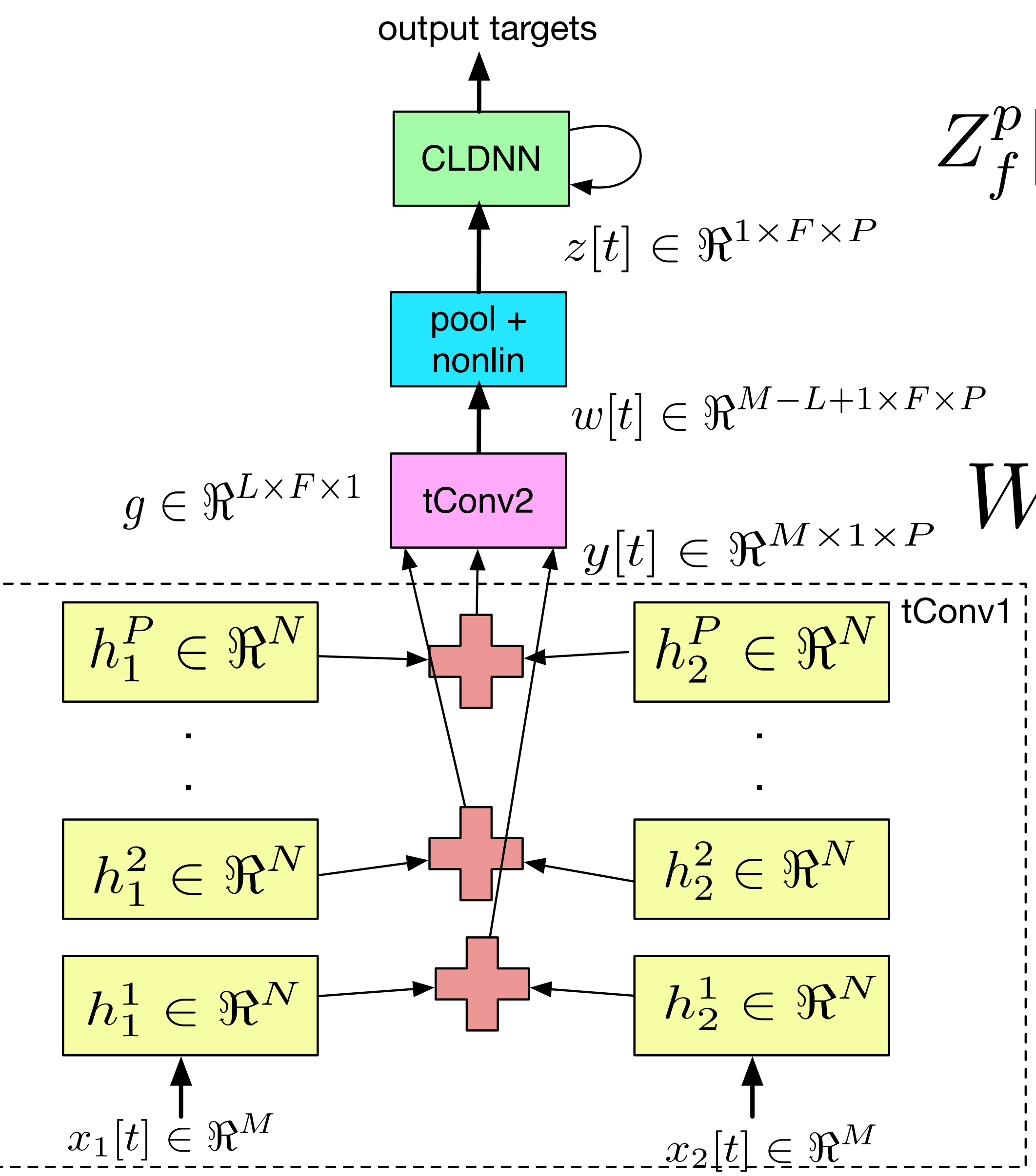
Time-Frequency Duality

- So far, all models have been formulated in the time domain

- Given the computational cost of a convolutional operator in time, the frequency dual of elementwise multiplication is of interest.

- Early layers of the network, to be phase sensitive use complex weights.

Factored Models in Frequency



Complex Linear Projection

$$Z_f^p[l] = \log \left| \sum_{k=1}^{N} W_f^p[l,k] \right|$$

$$W_f^p[l] = Y^p[l] \cdot G_f$$

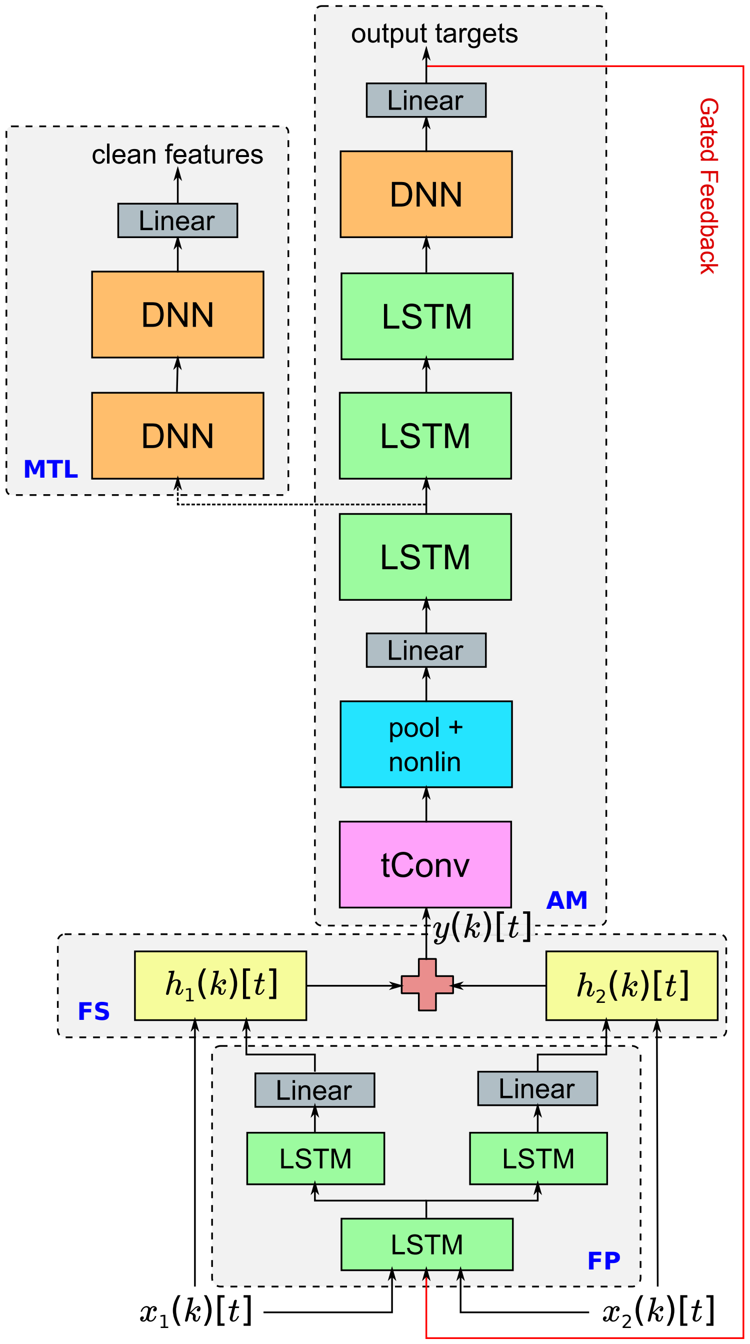
Linear Projection of Energy

$$Z_f^p[l] = G_f \times (\hat{Y}^p[l])^\alpha$$

$$\hat{Y}^p[l,k] = |Y^p[l,k]|^2$$

$$Y^p[l] = \sum_{c=1}^{C} X_c[l] \cdot H_c^p$$

Neural Adaptive Beamforming in Frequency



- The filter prediction LSTM computes two 257 length complex filter (4 x 257 weights >> 25 taps in the time domain)
- Filters are applied to the complex FFT input signals and summed
- The resulting representation is then input to a LDNN with either CLP or LPE akin to the factored model.

Frequency Model Performance

NAB

Model	WER CE	Parameters	Total M+A
Raw	20.5	24.6M	35.3M
NAB CLP	21.0	24.7M	25.1M

Factored

Model	Spatial M+A	Spectral M+A	Total M+A	WER Seq
CLP	10.3k	655.4k	19.6M	17.2
LPE	10.3k	165.1k	19.1M	17.2

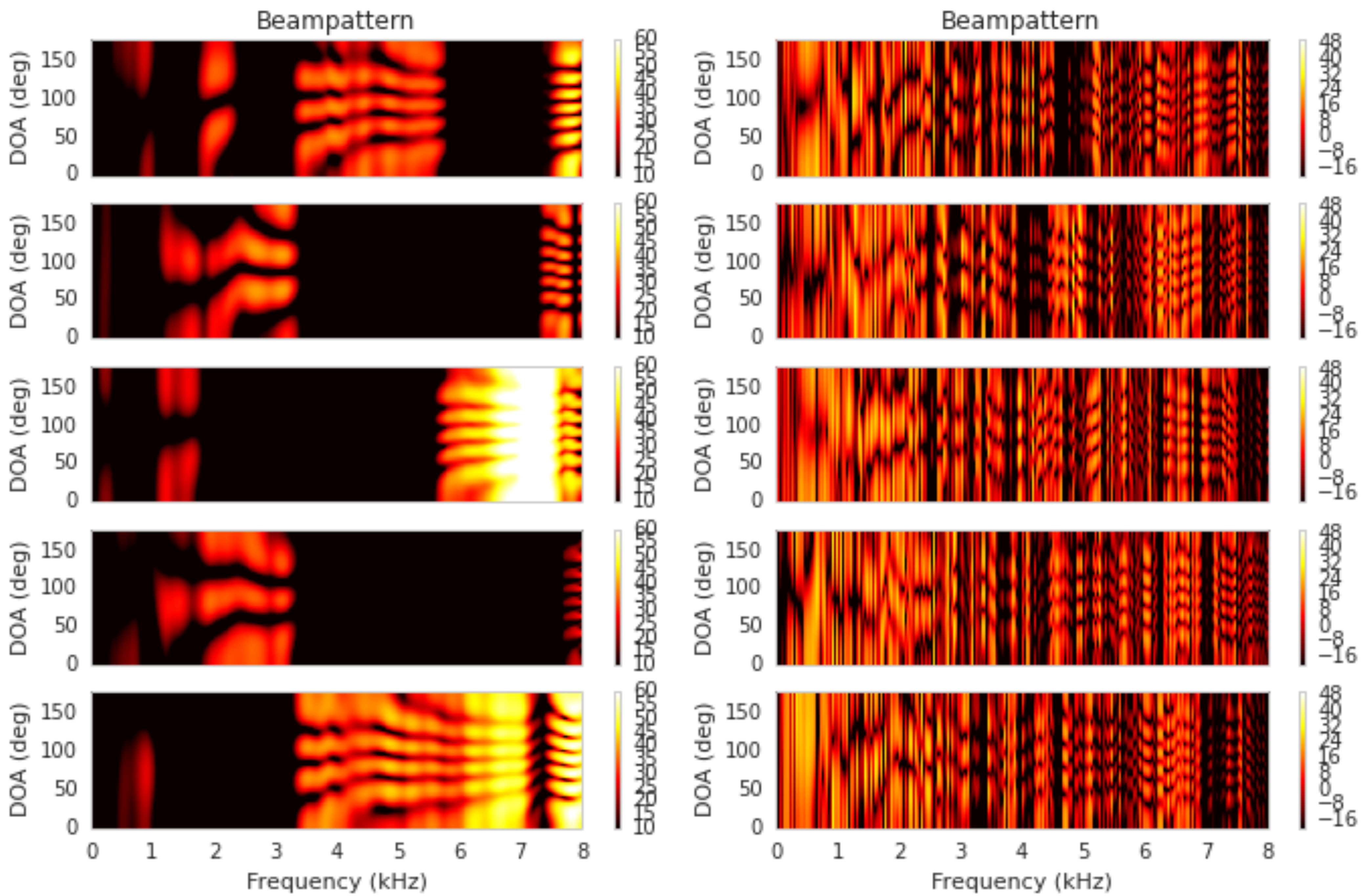
Factored increasing the model to 64ms/1024FFT

Model	Spatial M+A	Spectral M+A	Total M+A	WER Seq
Raw	906.1k	33.8M	53.6M	17.1
CLP	20.5k	1.3M	20.2M	17.1
LPE	20.5k	329k	19.3M	16.9

Time vs. Frequency Filters

(a) Factored model, time

(b) Factored model, frequency



Re-recorded Sets

- Two test sets from re-recording with the mic array “on the coffee table” or “on the TV stand”
- Only use 2-channel models as mic array configuration changed (circular vs. linear)

Model	Rev I	Rev II	Rev I Noisy	Rev II Noisy	Ave
1ch raw	18.6	18.5	27.8	26.7	22.9
2ch raw, unfactored	17.9	17.6	25.9	24.7	21.5
2ch raw, factored	17.1	16.9	24.6	24.2	20.7
2ch CLP, factored	17.4	16.8	25.2	23.5	20.7
2ch raw, NAB	17.8	18.1	27.1	26.1	22.3

Summary

- Google speech technology has really taken off with the “mobile revolution” together with the “neural network revolution”

- Novel applications like Google Home bring up new challenges and grounds research

- Neural network models appear attractive to incorporate several previously separate parts of the system: acoustic modeling + feature extraction + enhancement ….

  end-to-end modeling is a persistent direction

- Combining machine learning and “classical structures” provides an interesting framework for learning and comparing solutions.

Selected References

- H. Sak, A. Senior, and F. Beaufays, “Long Short-Term Memory Recurrent Neural Network Architectures for Large Scale Acoustic Modeling,” in Proc. Interspeech, 2014.
- T. N. Sainath, O. Vinyals, A. Senior, and H. Sak, “Convolutional, Long Short-Term Memory, Fully Connected Deep Neural Networks,” in Proc. ICASSP, 2015.
- Y. Hoshen, R. J. Weiss, and K. W. Wilson, “Speech Acoustic Modeling from Raw Multichannel Waveforms,” in Proc. ICASSP, 2015.
- T. N. Sainath, R. J. Weiss, K. W. Wilson, A. Senior, and O. Vinyals, “Learning the Speech Front-end with Raw Waveform CLDNNs,” in Proc. Interspeech, 2015.
- T. N. Sainath, R. J. Weiss, K. W. Wilson, A. Narayanan, M. Bacchiani, and A. Senior, “Speaker Localization and Microphone Spacing Invariant Acoustic Modeling from Raw Multichannel Waveforms,” in Proc. ASRU, 2015.
- T. N. Sainath, R. J. Weiss, K. W. Wilson, A. Narayanan, and M. Bacchiani, “Factored Spatial and Spectral Multichannel Raw Waveform CLDNNs,” in Proc. ICASSP, 2016.
- B. Li, T. N. Sainath, R. J. Weiss, K. W. Wilson, and M. Bacchiani, “Neural Network Adaptive Beamforming for Robust Multichannel Speech Recognition,” in Proc. Interspeech, 2016.
- Ehsan Variani, Tara N. Sainath, Izhak Shafran, Michiel Bacchiani “Complex Linear Projection (CLP): A Discriminative Approach to Joint Feature Extraction and Acoustic Modeling”, in Proc. Interspeech 2016
- Tara N. Sainath, Arun Narayanan, Ron J. Weiss, Ehsan Variani, Kevin W. Wilson, Michiel Bacchiani, Izhak Shafran, “Reducing the Computational Complexity of Multimicrophone Acoustic Models with Integrated Feature Extraction”, in Proc. Interspeech 2016
- T. N. Sainath, A. Narayanan, R. J. Weiss, K. W. Wilson, M. Bacchiani, and I. Shafran, “Improvements to Factorized Neural Network Multichannel Models,” in Proc. Interspeech, 2016.