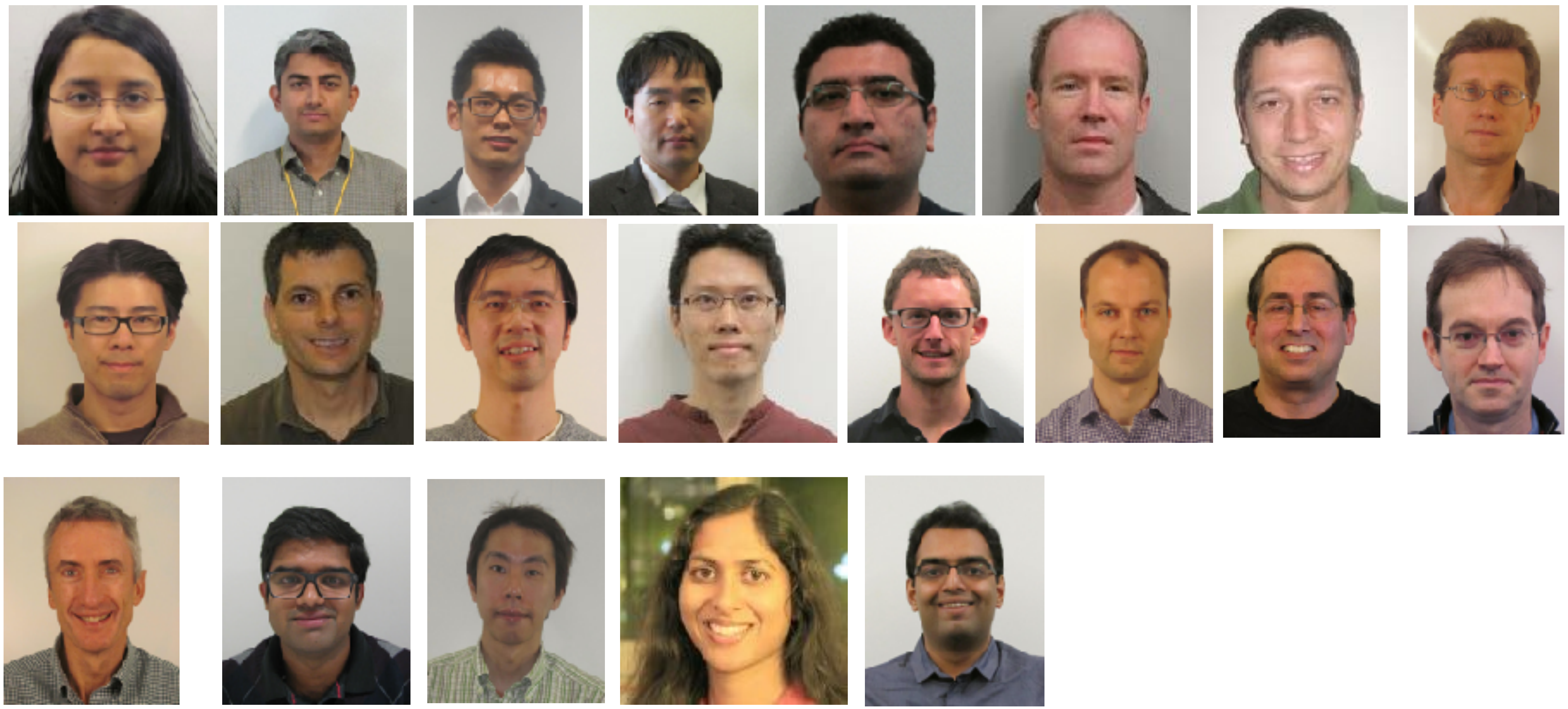
(Towards) next generation acoustic models for speech recognition

Erik McDermott
Google Inc.

It takes a village



… and 250 more colleagues in the Speech team

Overview

- The past: some recent history

- The present: the "conventional" state-of-the-art, from the perspective of Farfield / Google Home.

- The future is already here? End2End.

- Longer-term: Deep Generative approach?

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

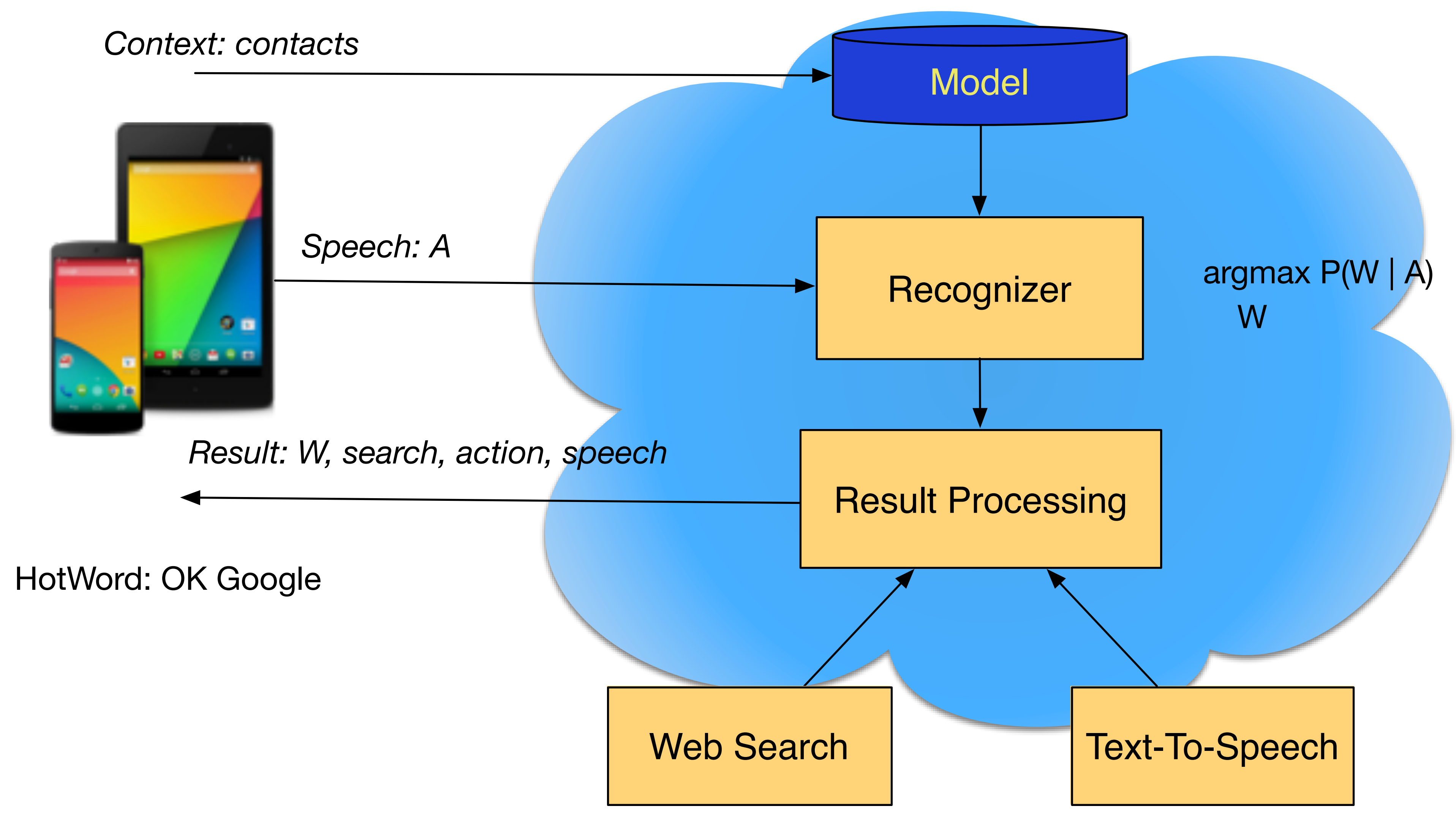
Launched early 2010

- 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 100+ languages and processes several centuries of speech each week

Mobile Application Overview



Recognition Models

Multi-lingual

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

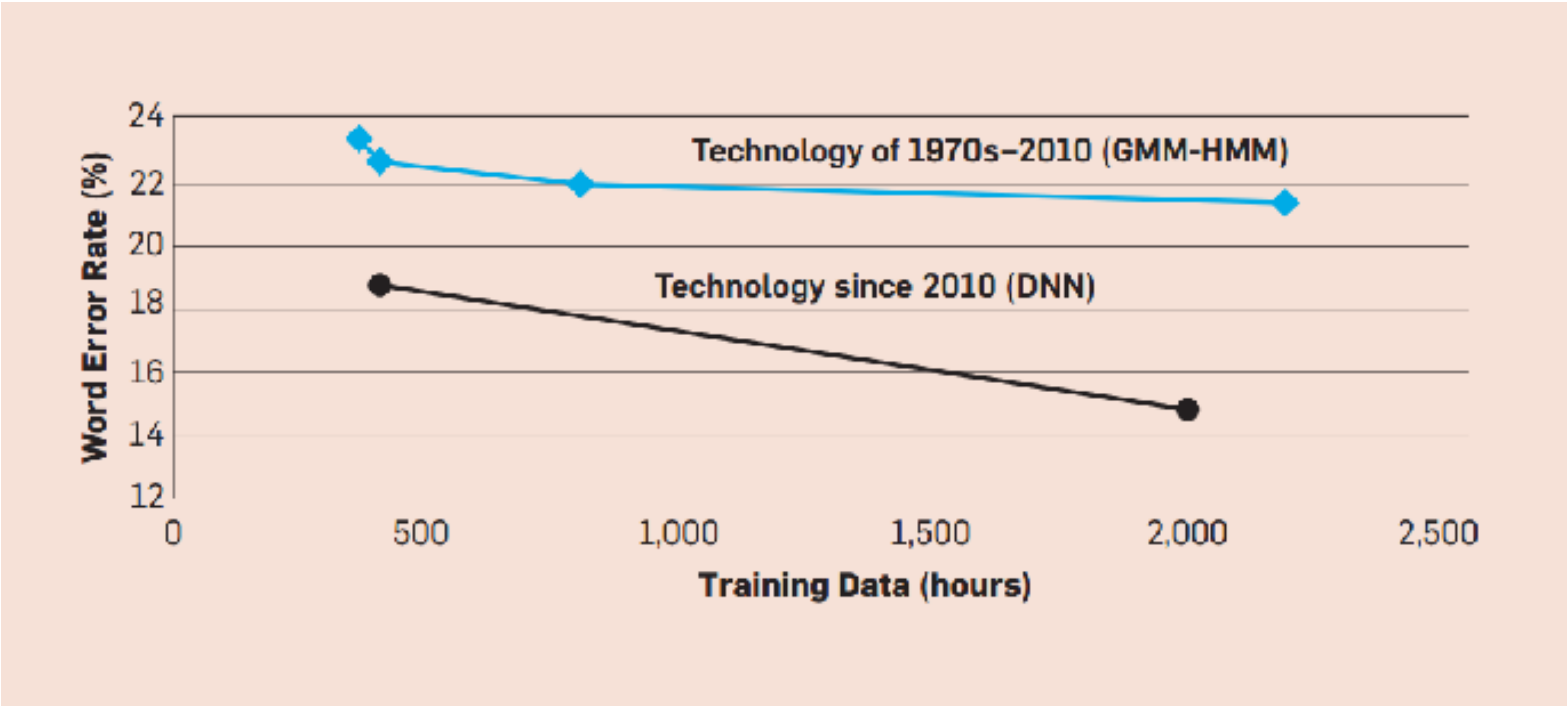
Finite State Transducers

Deep Neural Networks

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 Historical Perspective of Speech Recognition,"* Communications of the ACM, January 2014, Vol. 57, No 1.

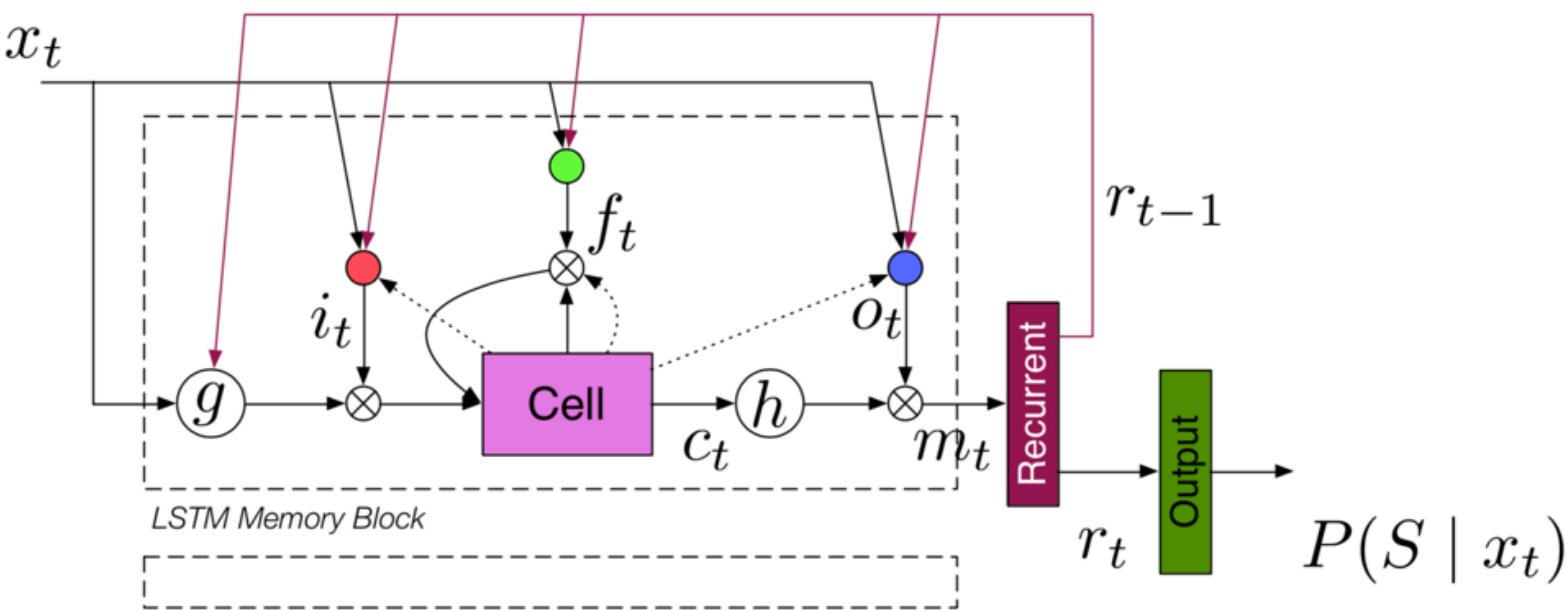
Accuracy Gains from Data and Modeling

- Initial results using DNNs in hybrid systems showed large gains (GMM 16.0% to DNN 12.2% with about 2k hours on VoiceSearch task)
- Additional gains from larger models
- Application of sequence models and sequence training

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

Long Short Term Memory

- Facilitates BPTT compared to vanilla RNNs.
- Trains efficiently.



Optimization with TensorFlow

- {CE,CTC} + {sMBR,WMBR}
- No observable differences between CE and CTC
- On-the-fly decoding for sMBR/WMBR on CPU driving LSTMs on GPU/TPU
- WMBR based on M. Shannon’s sampling-based approach (“EMBR”, Interspeech 2017).
- CTC can learn without alignments (FwdBkwd), but typically uses alignments as constraint for better latency.
- See “*End-to-end training of acoustic models for LVCSR with TensorFlow*”, Variani, Bagby, McDermott & Bacchiani, Interspeech 2017

Farfield

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

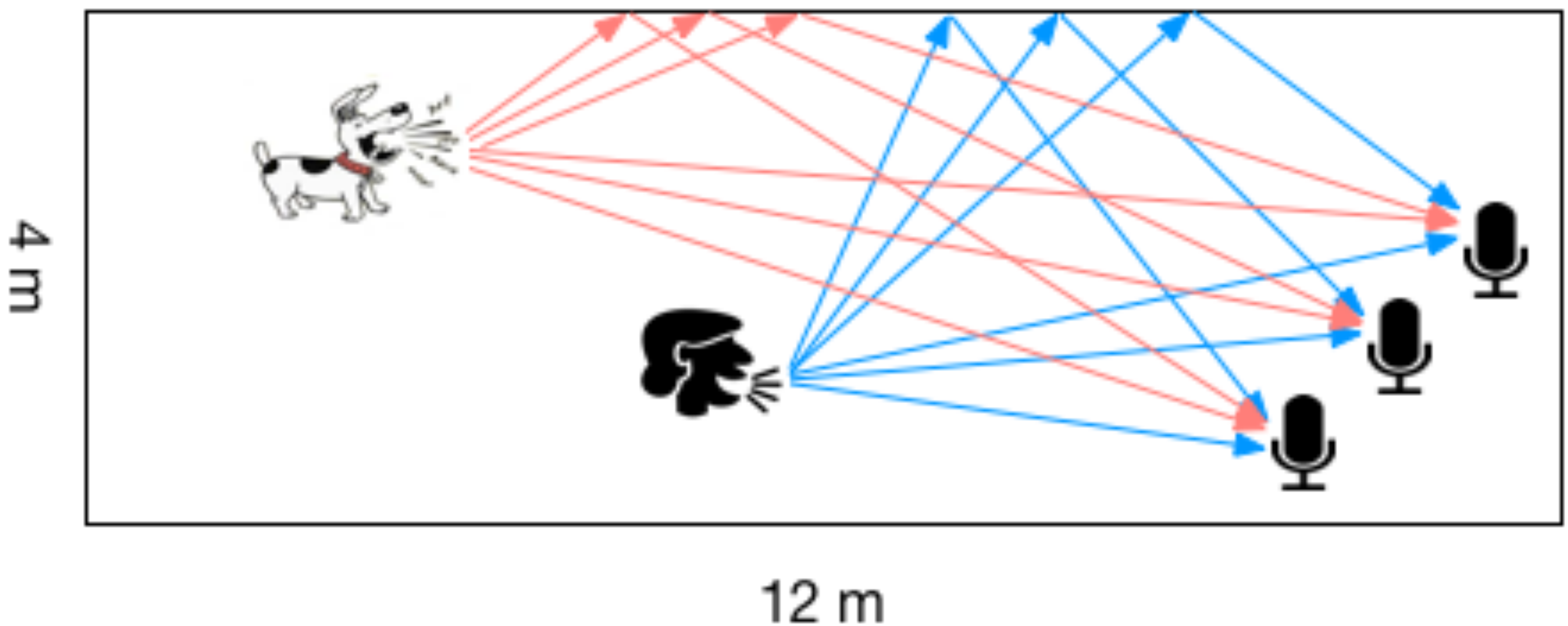
- Non-trivial engineering challenges: reverb, noise, level differences

Data Approach

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

Room Simulator

T60 = 500ms, SNR = 10dB



Study on Multi-channel processing with deep learning

- T. N. Sainath, R. J. Weiss, K. W. Wilson, B. Li, A. Narayanan, E. Variani, M. Bacchiani, I. Shafran, A. Senior, K. Chin, A. Misra and C. Kim "Multichannel Signal Processing with Deep Neural Networks for Automatic Speech Recognition," in IEEE Transactions on Speech and Language Processing, 2017.

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

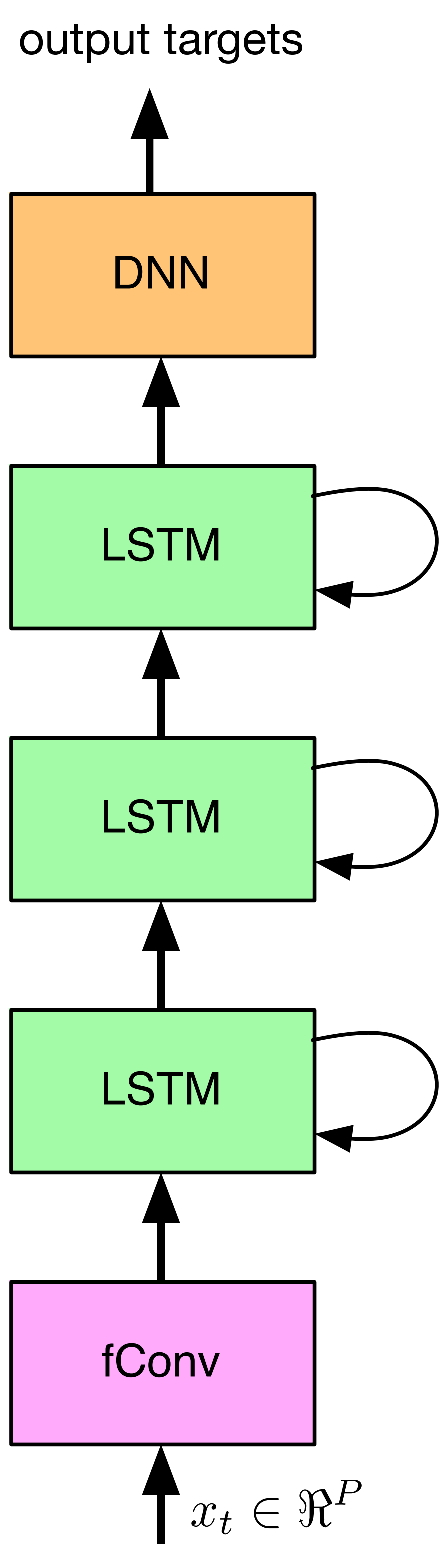
- Evaluate on a 30k voice search utterance set, about 20 hours
- One version simulated like the training set

- 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

baseline CLDNN



- Added accuracy improvements from combining layers of different types.

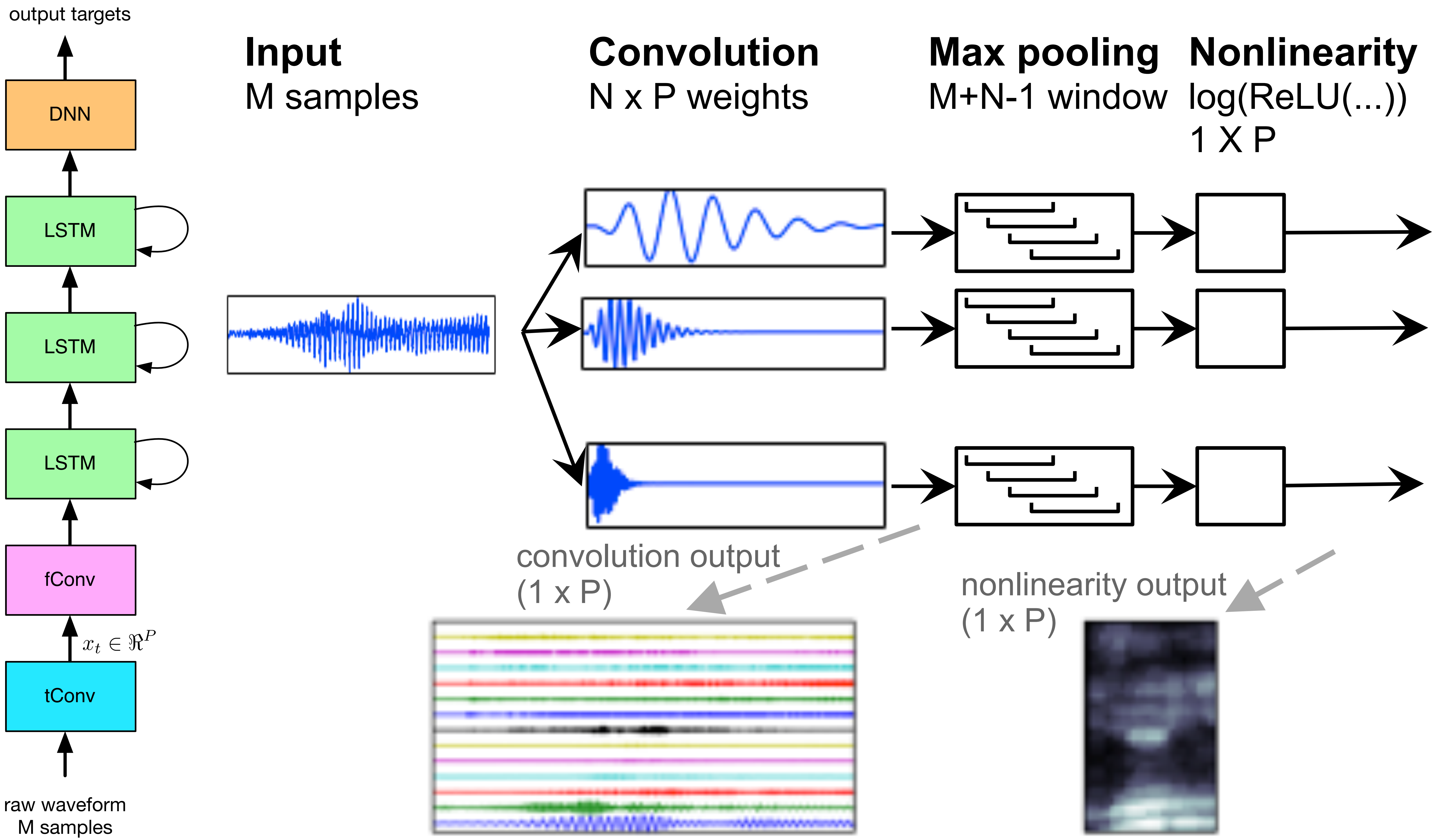
2000 hour clean training set,
20 hour clean test set

	CE	Sequence
LSTM	14.6	13.7
CLDNN	13.0	13.1

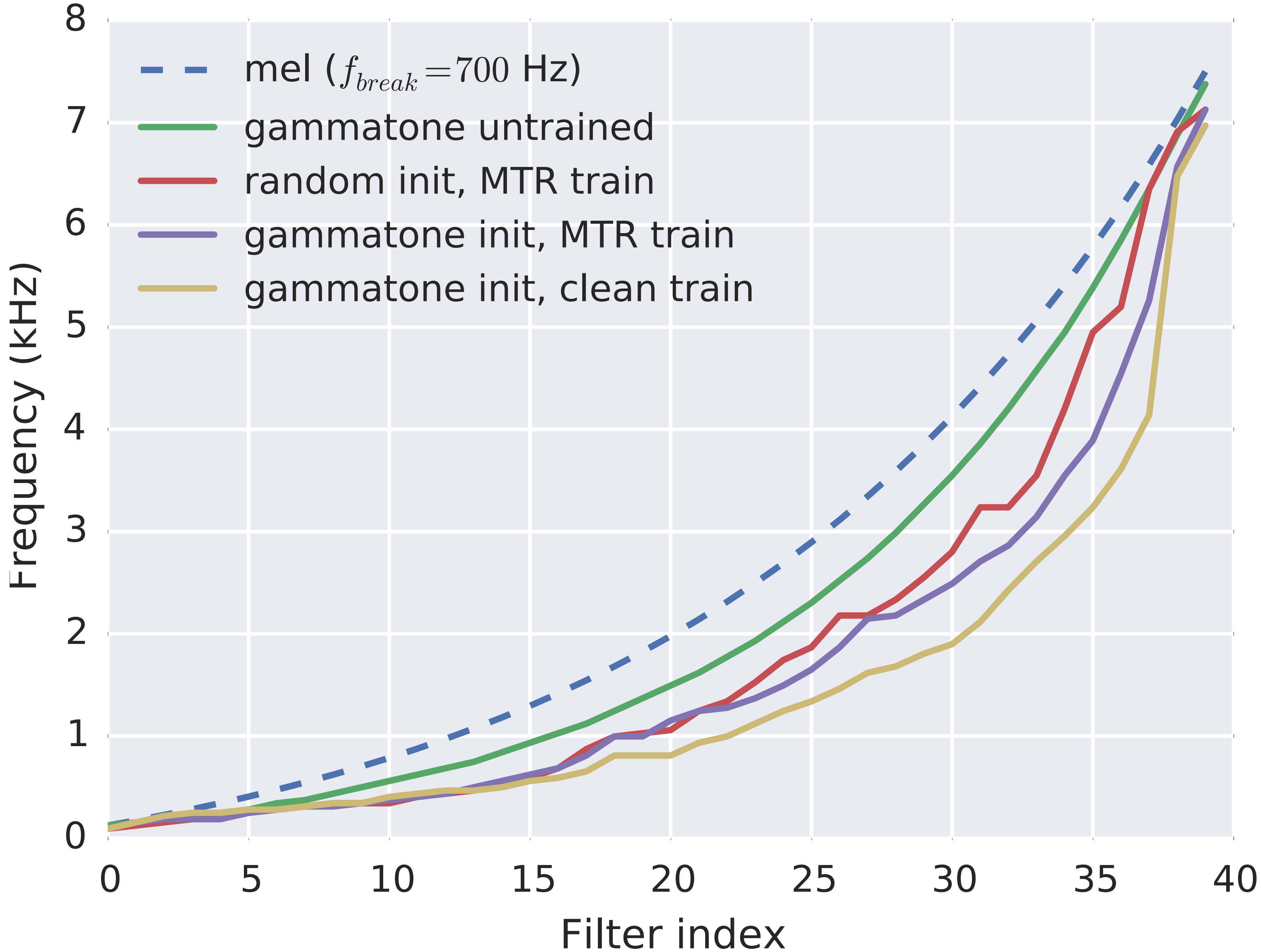
2000 hour MTR training set,
20 hour noisy test set

	CE	Sequence
LSTM	20.3	18.8
CLDNN	19.4	17.4

Raw Waveform Models



Raw Waveform Performance



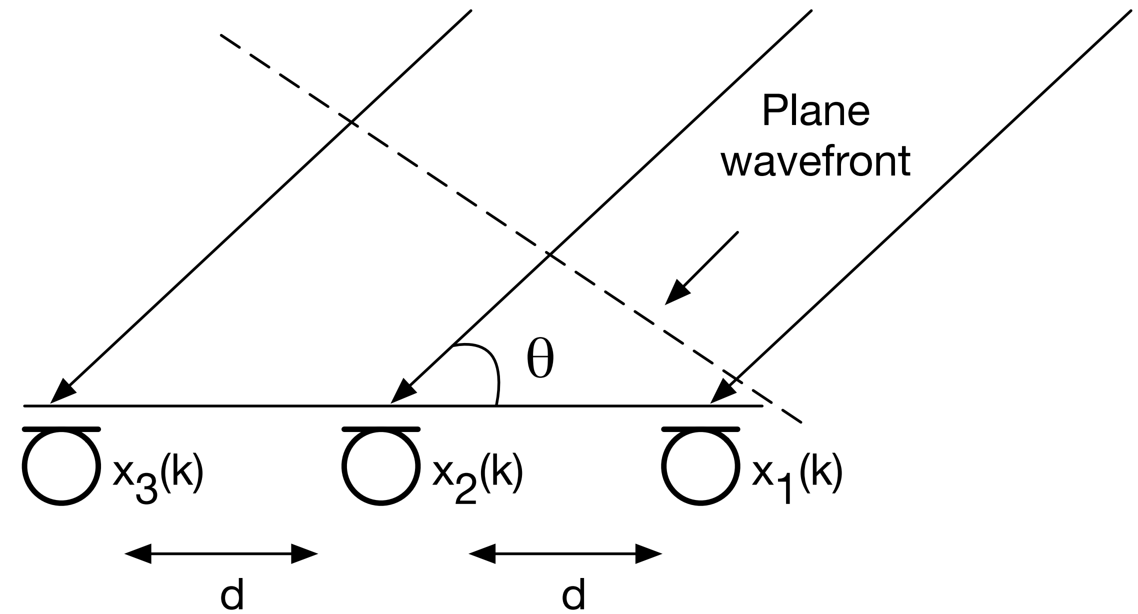
Model	Log Mel	Raw
C1L3D1	16.2	16.2
L3D1	16.5	16.5
D6	22.3	23.2

Multi-channel Enhancement

Localization

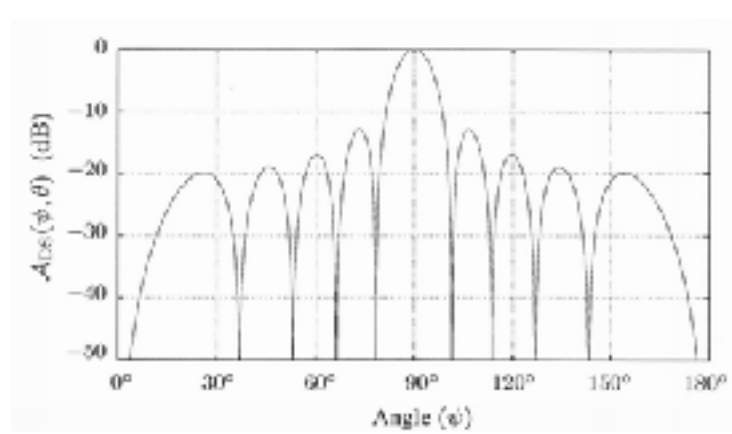
$$\tau_{ij} = \frac{d(i-j)\cos(\theta)}{c}$$

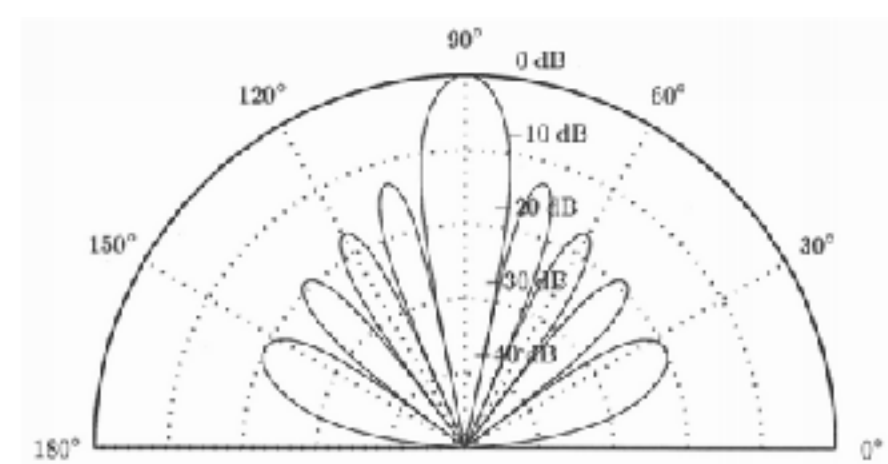
$$\hat{\tau_{ij}} = \underset{\tau}{\operatorname{argmax}} \sum_{t=0}^{L} x_i[t]x_k[t-\tau]$$



Delay-and-Sum Beamforming

$$y(t,\theta) = \frac{1}{M}\sum_{i} x_i[t-\tau_i(\theta)]$$





Multi-channel ASR

- Common approach separates enhancement and recognition

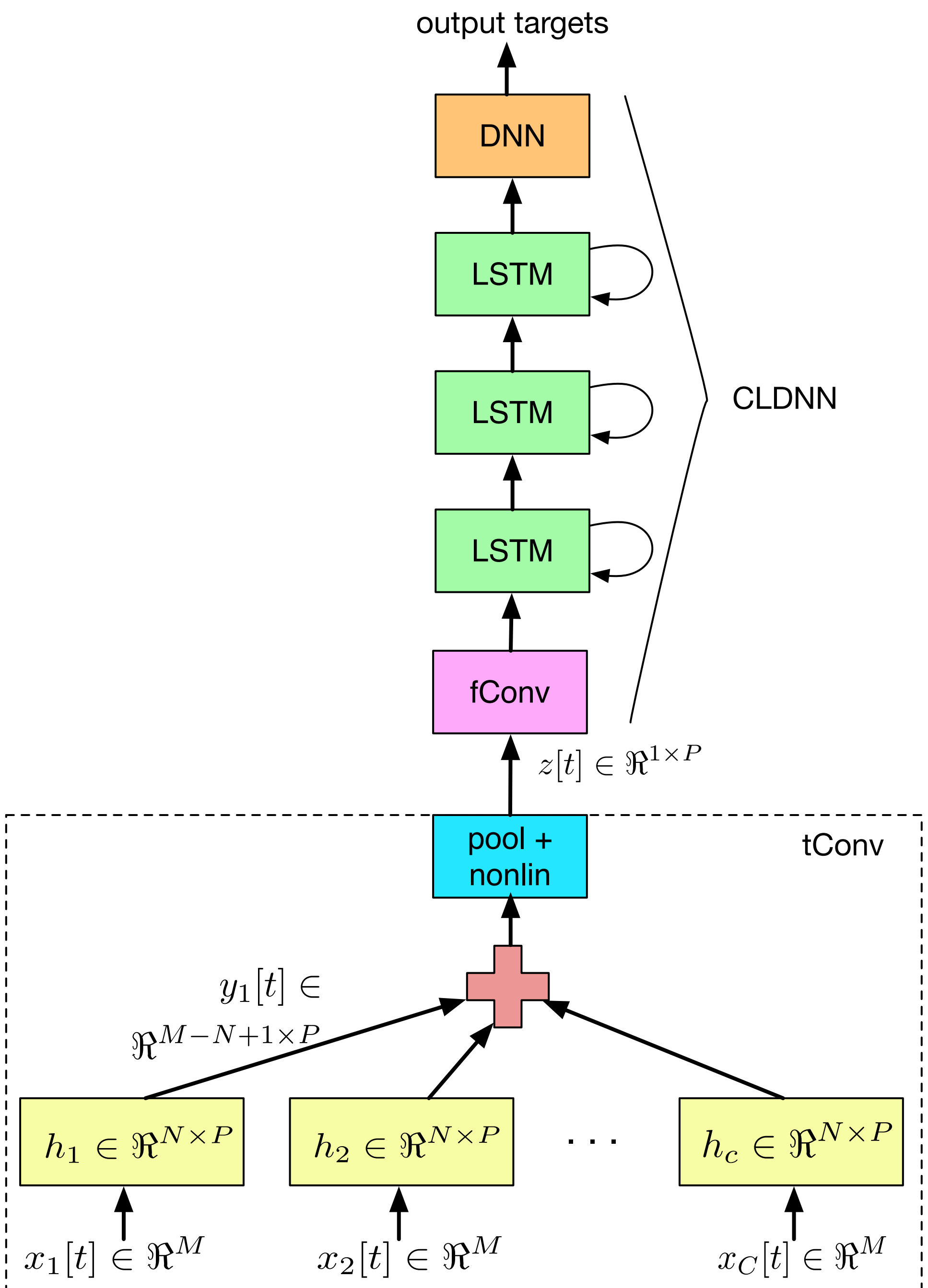
- Enhancement commonly done in localization, beamforming and postfiltering stages

- 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)

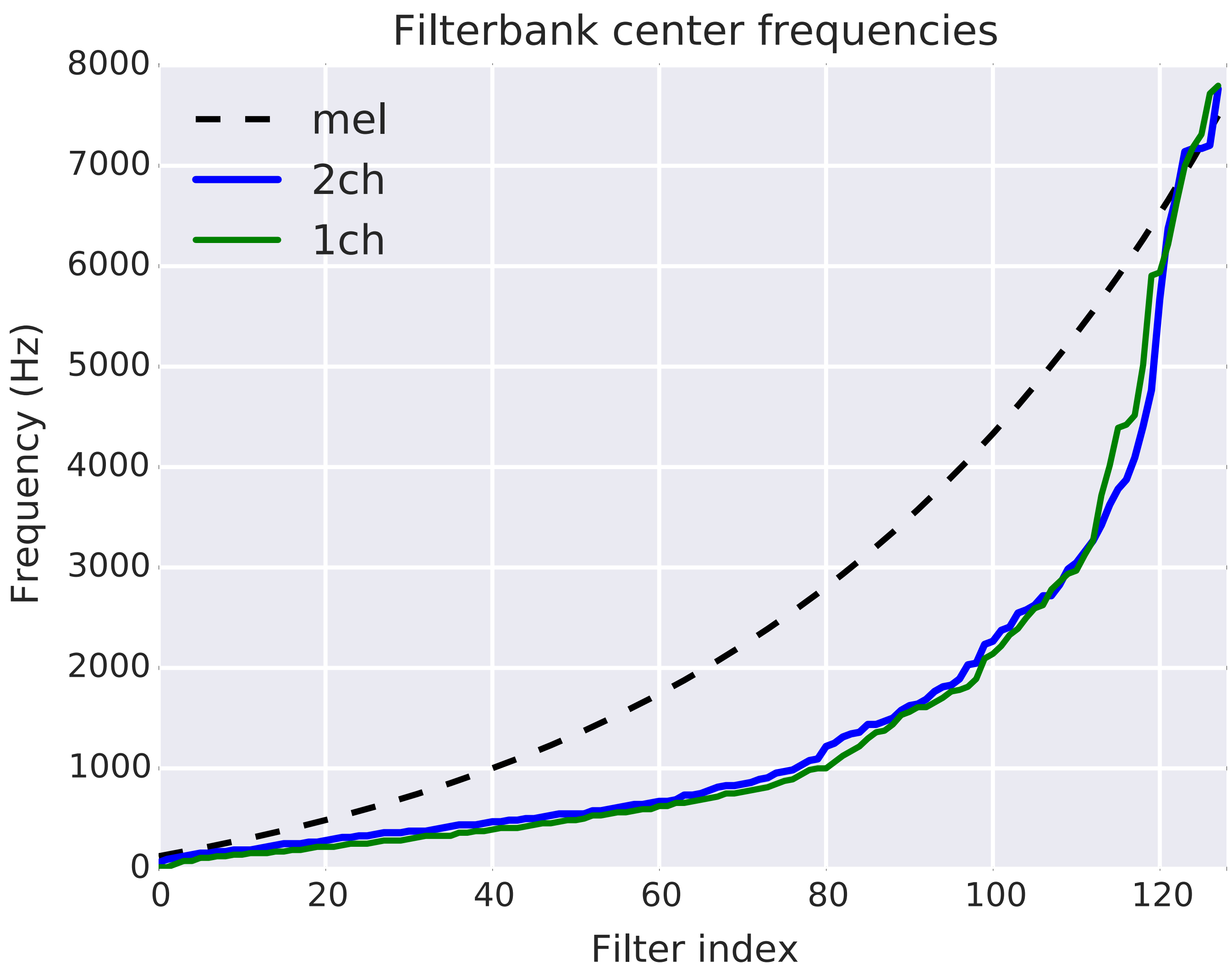
Raw Waveform & Multi-Channel



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

- Implicitly model steering delay with 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

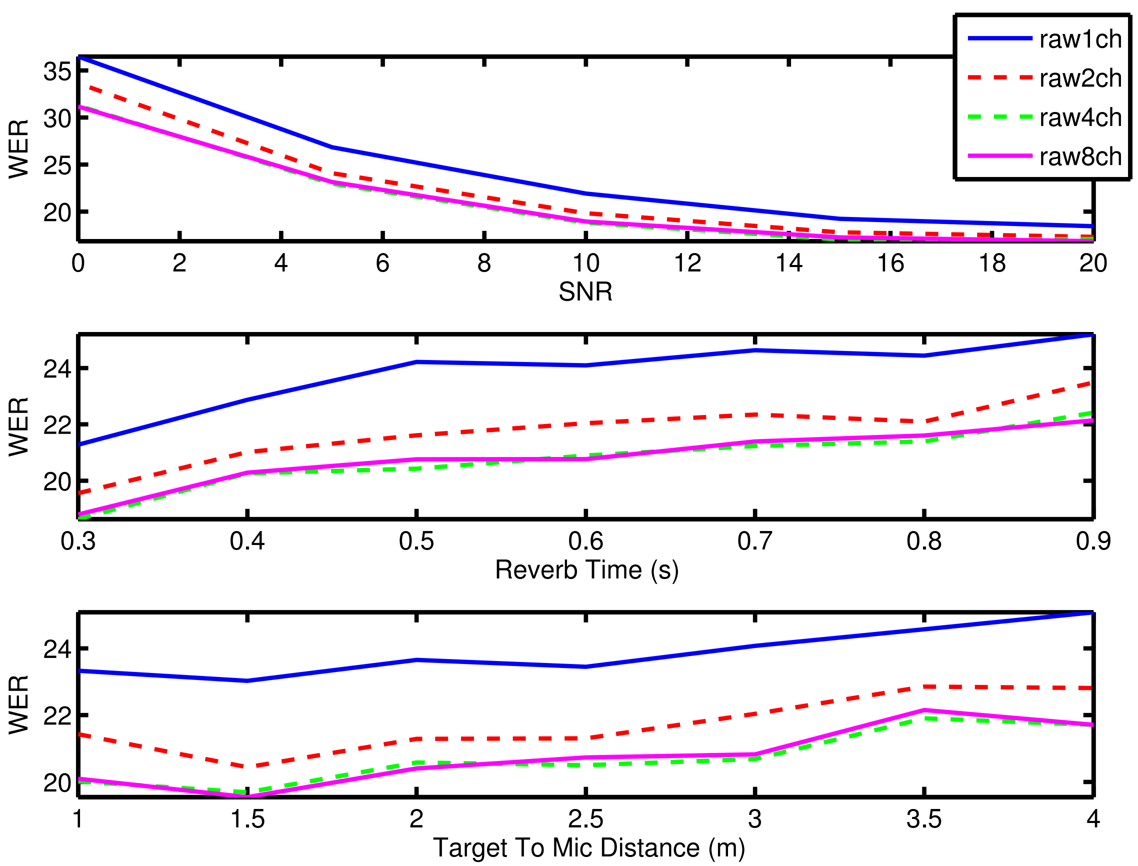
- The multi-channel raw waveform model does both beam forming as well as localization.

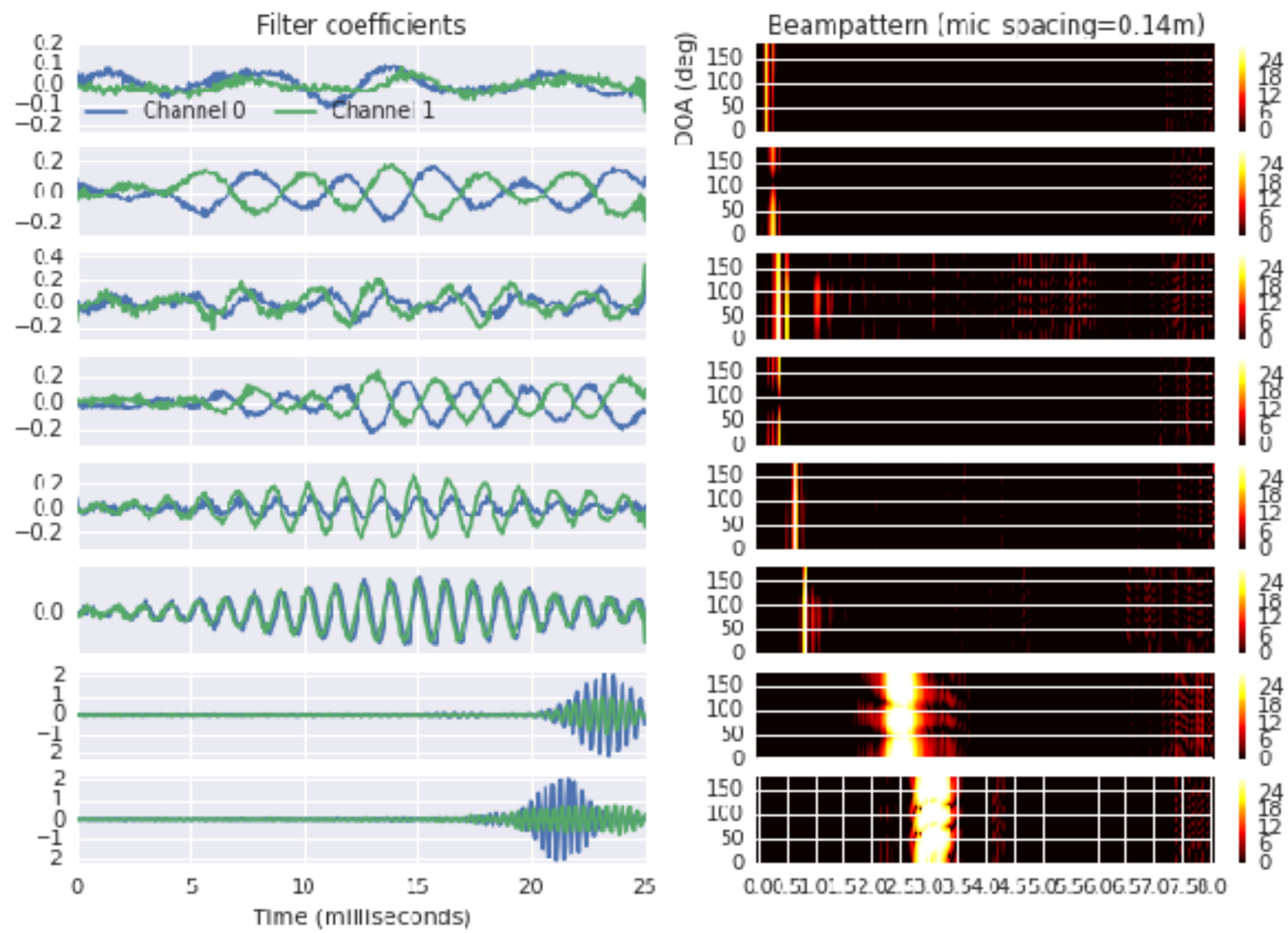
- Train a Delay-and-Sum (D+S) single channel signals with the oracle Time Delay of Arrival (TDOA)

- 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





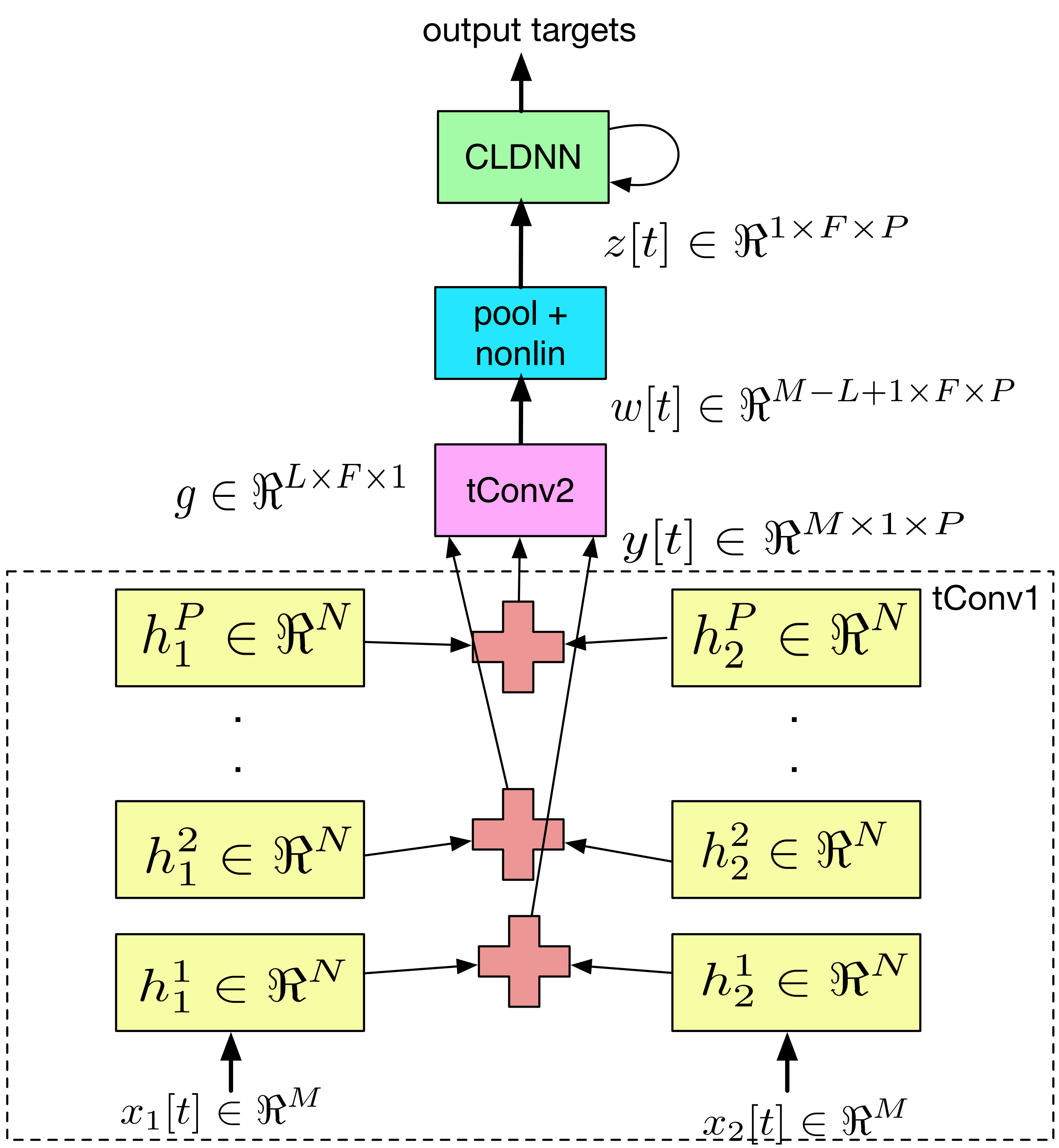
Multi-Channel Raw Waveform Summary

- 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

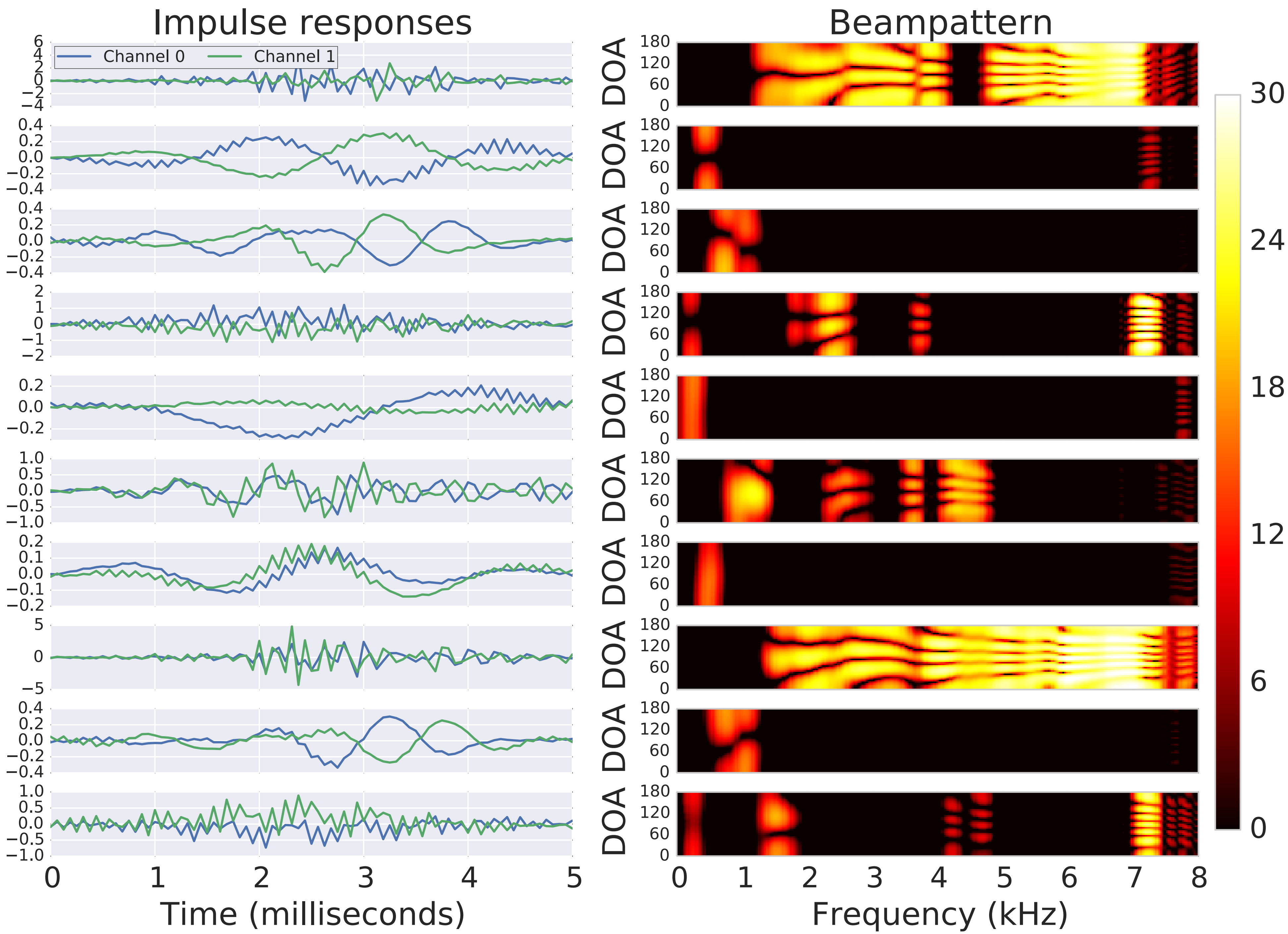
All systems 128 filters

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

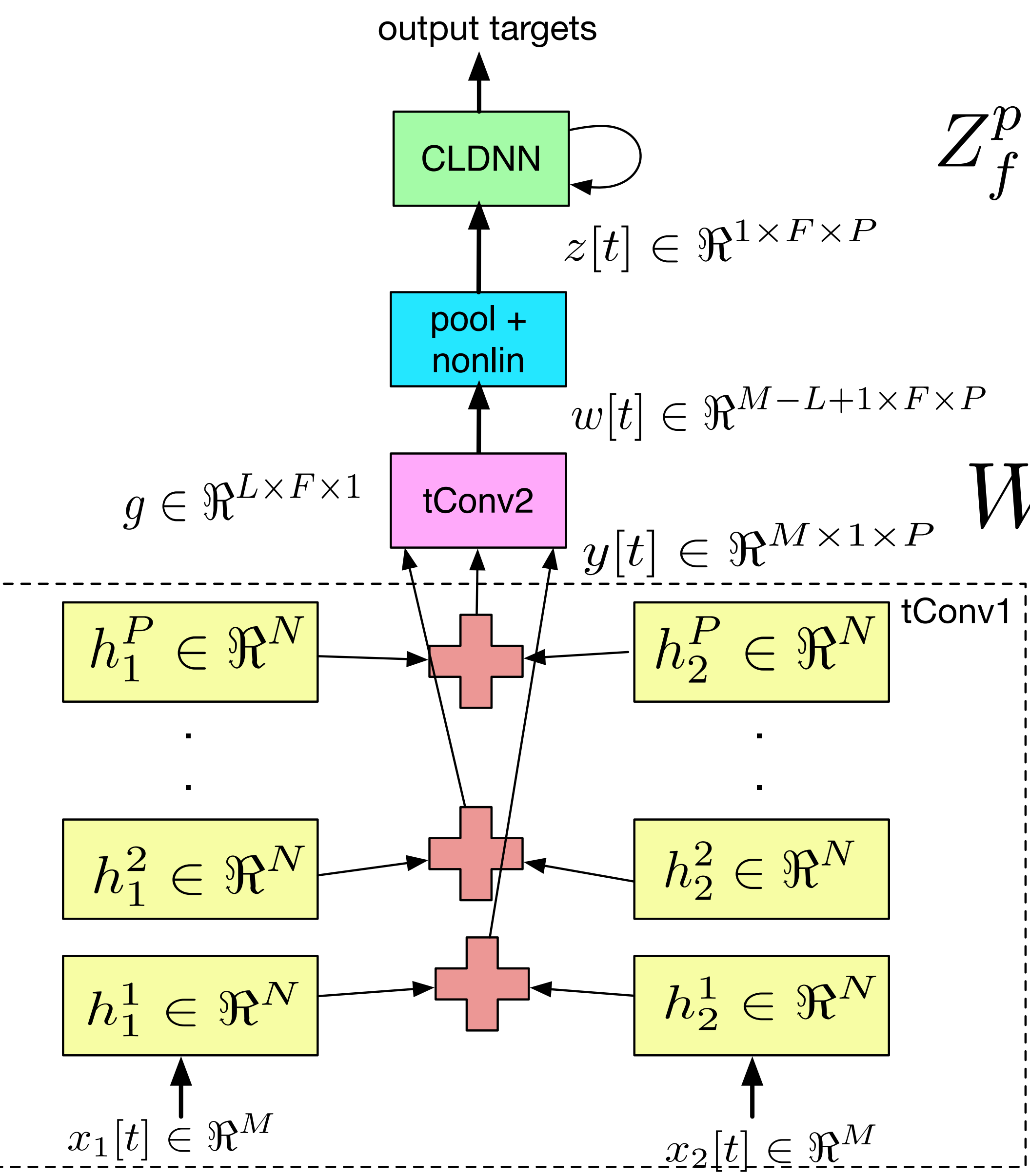
- Performance improvements remain after sequence training

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

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

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

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

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

Frequency Model Performance

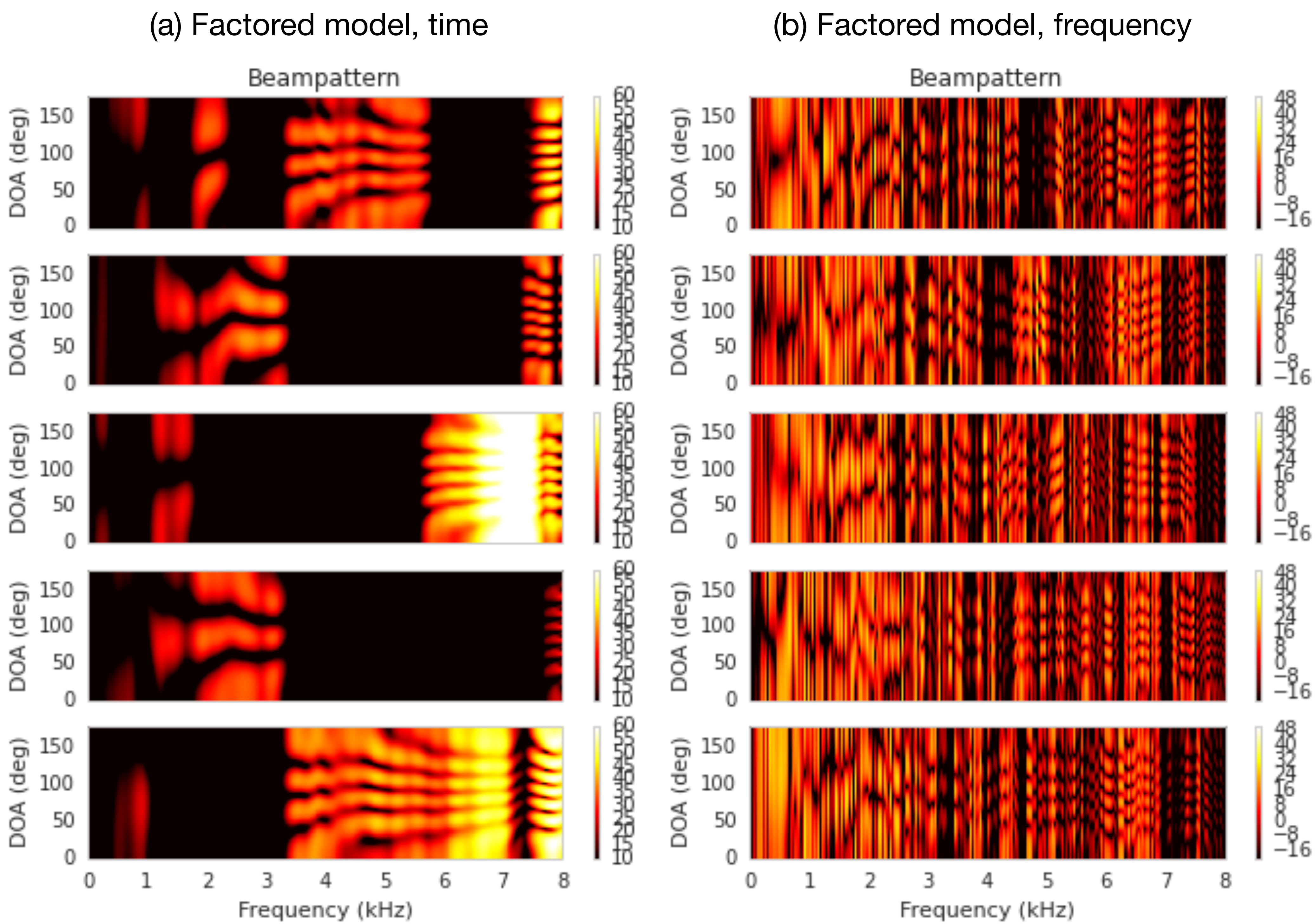
Factored

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



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

Google Home recent setup

- “Acoustic modeling for Google Home”, Li et al., Interspeech 2017
- 100 MTR room configurations → 4 million room configurations (Kim et al., Interspeech 2017)
- 2000 hours → 18,000 hours Voice Search training data
- Use of 4000 hours of Home real world traffic.
- Online Weighted Prediction Error (WPE) (based on Yoshioka & Nakatani)
- factored CLP; CLDNN → GridLSTM

Google Home recent results

WERs on Home eval set

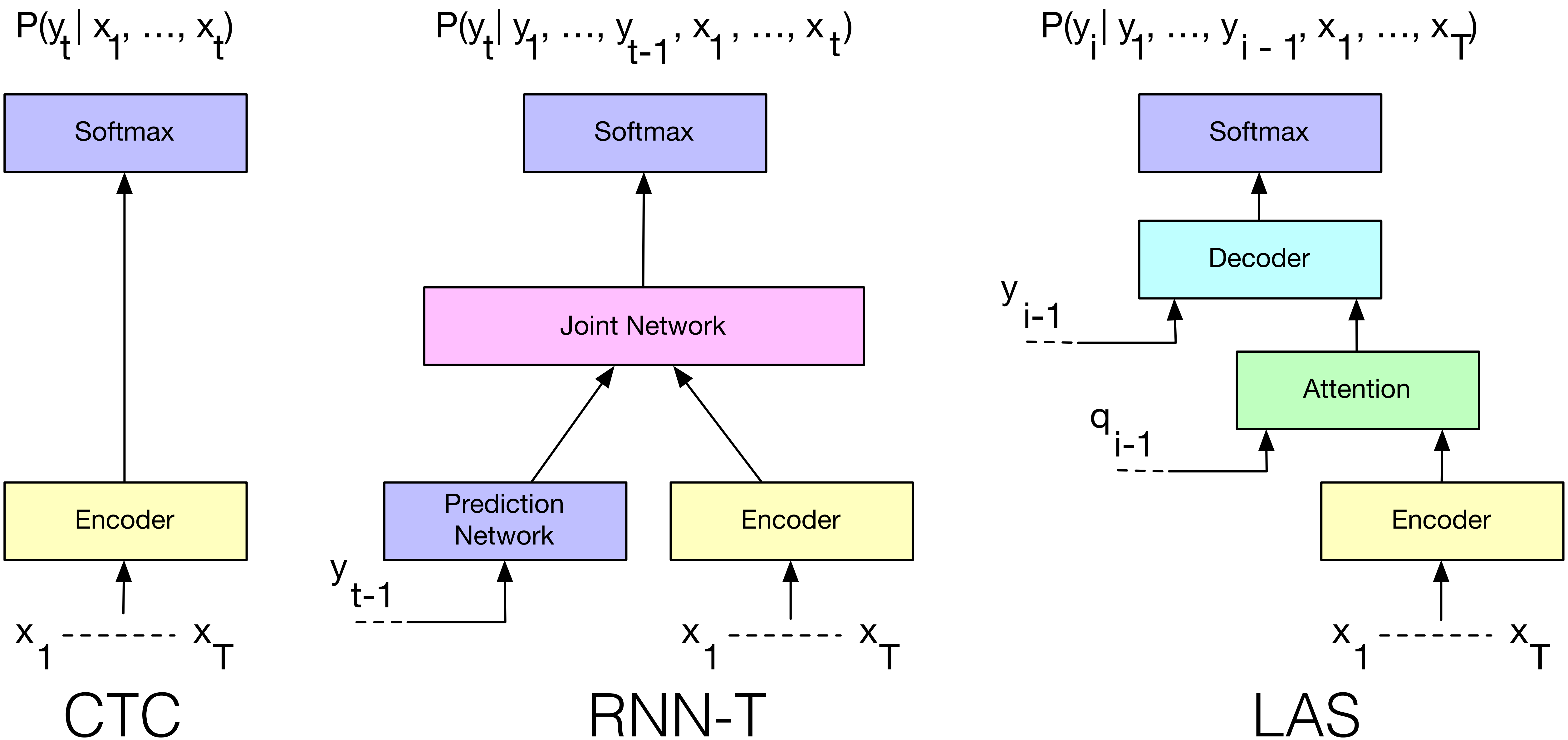
Model	Full	Clean	Noise Type		
			Speech	Music	Other
prod	6.1	5.1	8.5	6.2	6.0
home	5.1	4.9	6.3	5.1	5.0
home(adapt)	4.9	4.7	6.1	4.9	4.8

Most utterances are simple/low-perplexity:

- weather
- play XYZ
- change volume
- etc.

End-to-End Models

- Modeling string to string directly avoids any independence assumptions and allows joint optimization of the whole model.



Implications/Limitations

- **PROS**
  - Simplicity: no lexicon design, no tuning
  - No independence assumptions, joint optimization
- **CONS**
  - Need “complete data”; speech/text pairs
  - Not an online/streamable model
  - No clear input for manual design/“biasing”
  - Performance is poor on proper nouns / rare words.

The new state-of-the art?

- CC Chiu et al., “State-of-the-art speech recognition with sequence-to-sequence models”, Interspeech 2017.

- Reaching/surpassing results for standard hybrid model, e.g. CE + LSTM

- But issues with comparing results, details matter…

- .. and ongoing issues with streamability, LM biasing, rare words.

- Large number of topics to explore.

The path not (yet) taken: Waking up from the supervised, discriminative training dream?

- Is training on vast amounts of labelled training data really the future? Cost, freshness issues.
- Clearly a far vaster amount of unlabeled data is out there.
- Cf. Yan Le Cun’s plenary at ICASSP: use of predictive models, getting ground truth from the world.

ASR & TTS have grown closer, but are still quite distinct

- ASR: Limited generative models & discriminative training → Much richer discriminative models

[Though Hybrid Model fakes generative character at some level]

- TTS: Limited generative models → Much richer generative models

- How about a deep generative model for ASR?

Discriminative vs. generative models for ASR

- Discriminative "end-to-end" model, e.g. LAS

$$P(\mathrm{w}|\mathrm{x}) = \prod_{k} P(w_k|w_1, ..., w_{k-1}, A_k(\mathrm{x})) \tag{1}$$

- Combine with separate language model & sequence training:

$$Blend(\mathrm{x}, \mathrm{w}) = P(\mathrm{w}|\mathrm{x})^{\alpha} * P(\mathrm{w})^{1-\alpha} \tag{2}$$

- Cf. generative model:

$$p(\mathrm{x}, \mathrm{w}) = p(\mathrm{x}|\mathrm{w}) * P(\mathrm{w}) \tag{3}$$

$$P(\mathrm{w}) = \textstyle\prod_k P(w_k|w_1, ..., w_{k-1}) \tag{4}$$

$$p(\mathrm{x}|\mathrm{w}) = \prod_{t} p(x_t|x_1, ..., x_{t-1}, \mathrm{w}) \tag{5}$$

Deep generative model for TTS

- WaveNet (van den Oord et al. 2016):
  - Probability of a waveform (unconditioned):
  $$p(\mathrm{x}) = \prod_t p(x_t|x_1, ..., x_{t-1}), \tag{6}$$
  where observed samples x_t are targets of N-way quantized softmax trained with CE, using e.g. a DNN with dilated convolutions.
  - Conditional WaveNet:
  $$p(\mathrm{x}|\mathrm{h}) = \prod_t p(x_t|x_1, ..., x_{t-1}, \mathrm{h}), \tag{7}$$
  where the input h represents e.g. speaker and text info.
- Mixture density networks (Zen & Senior, 2014; Schuster 1997)
$$p(x_t|\mathrm{h}) = \sum w(x_{1:t-1}, \mathrm{h}) N(x_t|\mu(x_{1:t-1}, \mathrm{h}), \sigma(x_{1:t-1}, \mathrm{h})) \tag{8}$$

Deep generative model for ASR

- Define predictive, generative likelihood of observation feature vector x_t conditioned on all previous x_t and symbol sequence w:

$$p(\mathrm{x}|\mathrm{w}) = \prod_t p(x_t|x_1, ..., x_{t-1}, \mathrm{w}), \tag{9}$$

- Combine with LM for decoding & sequence training:

$$p(\mathrm{x}, \mathrm{w}) = p(\mathrm{x}|\mathrm{w}) * P(\mathrm{w}) \tag{10}$$

$$P(\mathrm{w}) = \textstyle\prod_k P(w_k|w_1, ..., w_{k-1}) \tag{11}$$

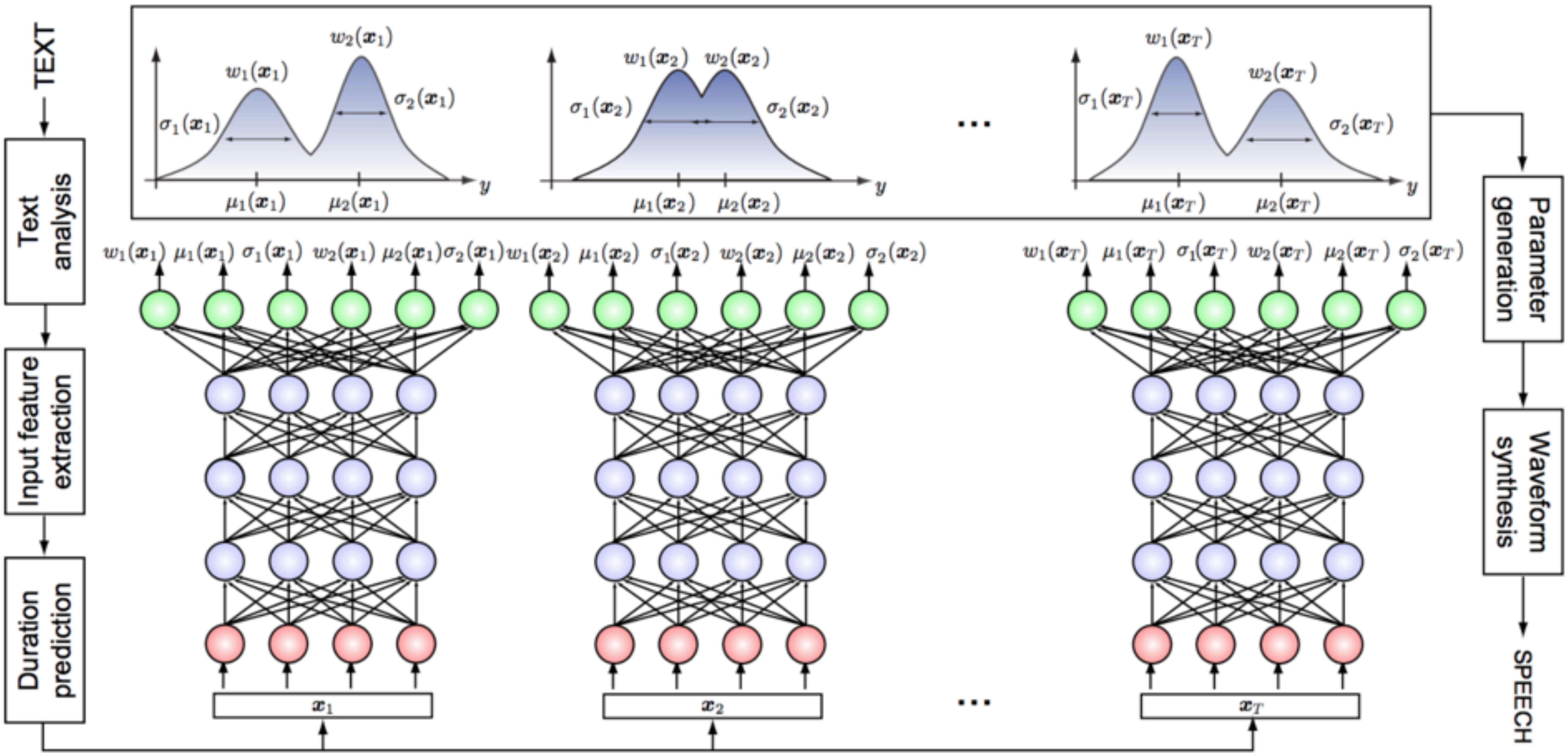
- Cf. hybrid model for LSTMs:

$$p(\mathrm{x}|\mathrm{w}) = \prod_t P(w_t|x_1, ..., x_t)/P(w_t) \tag{12}$$

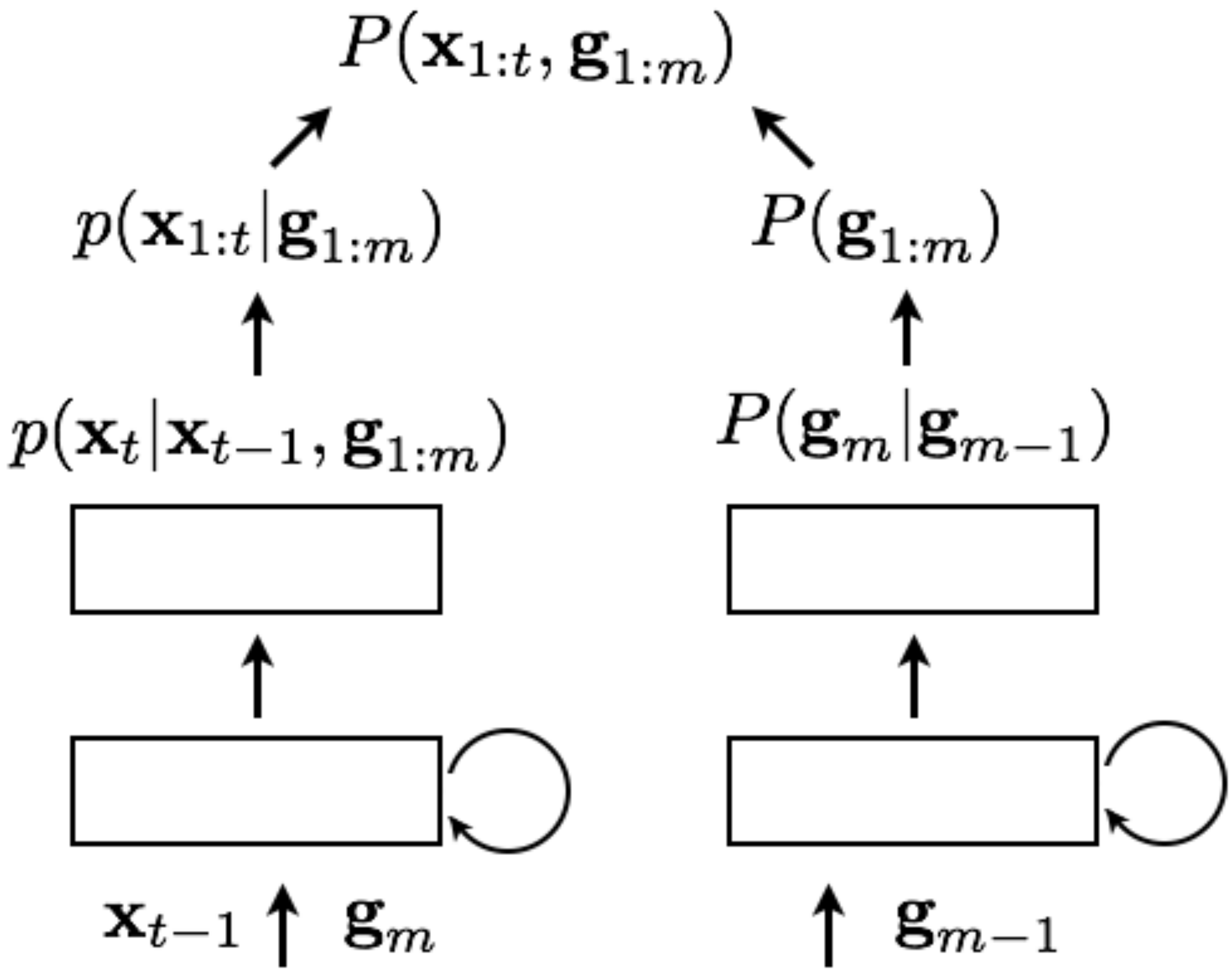
- Cf. ideal discriminative model

$$P(\mathrm{w}|\mathrm{x}) = \prod_k P(w_k|w_1, ..., w_{k-1}, x_1, ..., x_T) \tag{13}$$

Deep Mixture Density Nets for TTS, Zen & Senior, 2014



RNN Generative Transducer



Speech Remains Exciting

- Speech technology is becoming remarkably mainstream
- Many opportunities and research questions remain to be answered to make it truly ubiquitous: devices, languages, people, applications
- Thinking is not dead: model structure vs. parameter optimization
- Wide adoption means large data opening a very large opportunity for research using machine learning

Selected References

- E. Variani, T. Bagby, E. McDermott & M. Bacchiani, "*End-to-end training of acoustic models for LVCSR with TensorFlow*", in Proc. Interspeech, 2017.
- M. Shannon, "Optimizing expected word error rate via sampling for speech recognition", in Proc. Interspeech, 2017.
- C. Kim et al., "Generation of large-scale simulated utterances in virtual rooms to train deep-neural networks for far-field speech recognition in Google Home", in Proc. Interspeech 2017.
- B. Li, T. Sainath, A. Narayanan, J. Caroselli, M. Bacchiani, A. Misra, I. Shafran, H. Sak, G. Pundak, K. Chin, K.-C. Sim, R. J. Weiss, K. W. Wilson, E. Variani, C. Kim, O. Siohan, M. Weintraub, E. McDermott, R. Rose, M. Shannon, "Acoustic modeling for Google Home", in Proc. Interspeech 2017.
- C.-C. Chiu et al., "State-of-the-art speech recognition with sequence-to-sequence models", in Proc. ICASSP 2018
- R. Prabhavalkar et al., "Minimum word error rate training for attention-based sequence-to-sequence models", in Proc. ICASSP 2018

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

Selected References

- T. N. Sainath, A. Narayanan, R. J. Weiss, E. Variani, K. W. Wilson, M, Bacchiani, I. 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.
- T. N. Sainath, R. J. Weiss, K. W. Wilson, B. Li, A. Narayanan, E. Variani, M. Bacchiani, I. Shafran, A. Senior, K. Chin, A. Misra and C. Kim "Multichannel Signal Processing with Deep Neural Networks for Automatic Speech Recognition," in IEEE Transactions on Speech and Language Processing, 2017.
- C. Kim, A. Misra, K. Chin, T. Hughes, A. Narayanan, T. N. Sainath and M. Bacchiani, "Generation of Simulated Utterances in Virtual Rooms to Train Deep Neural Networks for Far-field Speech Recognition in Google Home," in Proc. Interspeech, 2017.
- B. Li, T. N. Sainath, J. Caroselli, A. Narayanan, M. Bacchiani, A. Misra, I. Shafran, H. Sak, G. Pundak, K. Chin, K. Sim, R. J. Weiss, K. W. Wilson, E. Variani, C. Kim, O. Siohan, M. Weintraub, E. McDermott, R. Rose and M. Shannon, "Acoustic Modeling for Google Home," in Proc. Interspeech, 2017.
- R. Prabhavalkar, K. Rao, T. N. Sainath, B. Li, L. Johnson and N. Jaitly, "A Comparison of Sequence-to-Sequence Models for Speech Recognition," in Proc. Interspeech, 2017.
- R. Prabhavalkar, T. N. Sainath, B. Li, K. Rao and N. Jaitly, "An Analysis of "Attention" in Sequence-to-Sequence Models," in Proc. Interspeech, 2017.
- C. Chiu, T. N. Sainath, Y. Wu, R. Prabhavalkar, P. Nguyen, Z. Chen, A. Kannan, R. J. Weiss, K. Rao, K. Gonina, N. Jaitly, B. Li, J. Chorowski, M. Bacchiani, “State-of-the-Art Speech Reconition with Sequence-to-Sequence Models,” submitted to ICASSP, 2018
- A. Kannan, Y. Wu, P. Nguyen, T. N. Sainath, Z. Chen, R. Prabhavalkar, “An Analysis of Incorporating an External Language Model into a Sequence-to-Sequence Model,” submitted to ICASSP 2018
- T. N. Sainath, C. Chiu, R. Prabhavalkar, A. Kannan, Y. Wu, P. Nguyen, Z. Chen, “Improving the Performance of Online Neural Transducer Models,” submitted to ICASSP 2018
- R. Prabhavalkar T. N. Sainath Y. Wu P. Nguyen Z. Chen C. Chiu A. Kannan, "Minimum Word Error Rate Training for Attention-based Sequence-to-Sequence Models,” submitted to ICASSP 2018
- B. Li, T. N. Sainath, K. C. Sim, M. Bacchiani, E. Weinstein, P. Nguyen, Z. Chen, Y. Wu, K. Rao, “Multi-Dialect Speech Recognition with a Single Sequence-to-Sequence Model,” submitted to ICASSP 2018