Multichannel Raw-Waveform Neural Network Acoustic Models

Tara N. Sainath
December 17, 2017

(in collaboration with Ron J. Weiss, Kevin W. Wilson, Bo Li, Arun Narayanan, Michiel Bacchiani, Joe Caroselli, Matt Shannon, Golan Pundak, Ehsan Variani, Chanwoo Kim, Ananya Misra, Kean Chin, Izhak Shafran, Andrew Senior)

ASRU 2017

Agenda

- Motivation
- Neural Beamforming Architectures
  - Unfactored raw-waveform - uRaw
  - Factored raw-waveform - fRaw
  - Factored Complex Linear Prediction - fCLP
  - Neural Adaptive Beamforming - NAB
- Experimental Evaluations on More Realistic Data
- Conclusions

Motivation

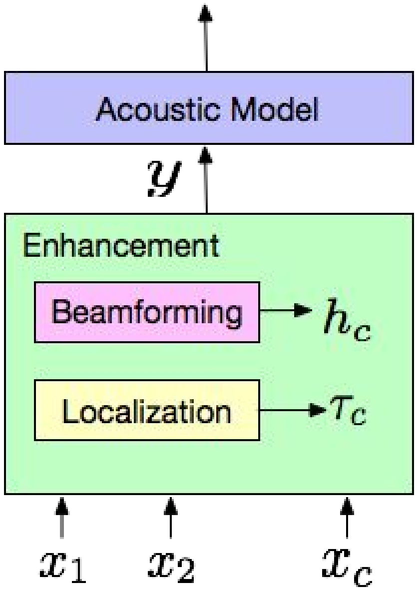
- Farfield speech recognition is becoming a new way to interact with devices at home.
- Farfield speech is difficult due to both **additive and reverberant noises.**
- Multi-channel signal processing techniques attempt to enhance signal and suppress noise.
- In this work, we detail different research ideas explored towards developing **Google Home**.

Typical Multi-channel Processing

- Most multichannel ASR systems use two **separate modules**
  1) Speech-enhancement (i.e., localization, beamforming)
  2) Single-channel acoustic model
- Traditional Filter+Sum (F+S) for **enhancement**

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

- Can we do enhancement and acoustic modeling **jointly**?



Neural-Beamforming Layers Explored in This Work

- We explore training a **neural beamforming layer** jointly with the acoustic model, using the raw-waveform to model fine time structure
- Traditional F+S
  - Learns localization τ_c for every utterance
  - Learns a filter h_c for every utterance

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

Neural Beamforming Architecture	Learning Methodology
Unfactored raw-waveform - uRaw	Time-domain filter h_c fixed after training
Factored raw-waveform - fRaw	Set of p time-domain filters h_c fixed after training
Factored Complex Linear Prediction - fCLP	Set of p frequency-domain filters h_c fixed after training
Neural Adaptive Beamforming - NAB	Time/frequency filter h_c updated at every time frame t

Related Work, Joint Multi-channnel Enhancement + AM

- [Seltzer, 2004] explored joint enhancement + acoustic modeling using a model-based GMM approach
- Beamformer with filter-based estimation network [Xiao, 2016]
  - Similar to the NAB model we will discuss [B. Li, 2016]
- Beamformer with mask estimation network [Heymann 2016, Erdogan 2016]
- Beamformer with both mask + filter estimation, end-to-end framework [Ochiai 2017]

Focus of our work is to detail the architectures explored for **Google HOME**.

Initial Experimental Setup

Training data:

- 3M English utterances
- 2,000 hours noisy data
- artificially corrupted with music, ambient noise, recordings of "daily life" environments
- SNRs: 0 ~ 30dB, avg. = 11dB
- Reverberation RT60: 0 ~ 900ms, avg. = 500ms
- 8 channel linear mic with spacing of 2cm
- Noise and speaker locations change per utt

Testing data:

- 13K English utterances
- 15 hours data
- simulated: matching training data
- Channel details:
    - 2 channel (1, 8): 14cm spacing
    - 4 channel (1, 3, 6, 8): 4-6-4cm spacing
    - 8 channel: 2cm spacing

Experiments are conducted to understand benefit of each proposed method.

Unfactored Raw-Waveform Model

T. N. Sainath, R. J. Weiss, K. W. Wilson, A. Narayanan, M. Bacchiani and A. Senior, "Speaker Location and Microphone Spacing Invariant Acoustic Modeling from Raw Multichannel Waveforms," in Proc. ASRU, December 2015.

Motivation from Traditional Filter + Sum

- Traditional filter + sum

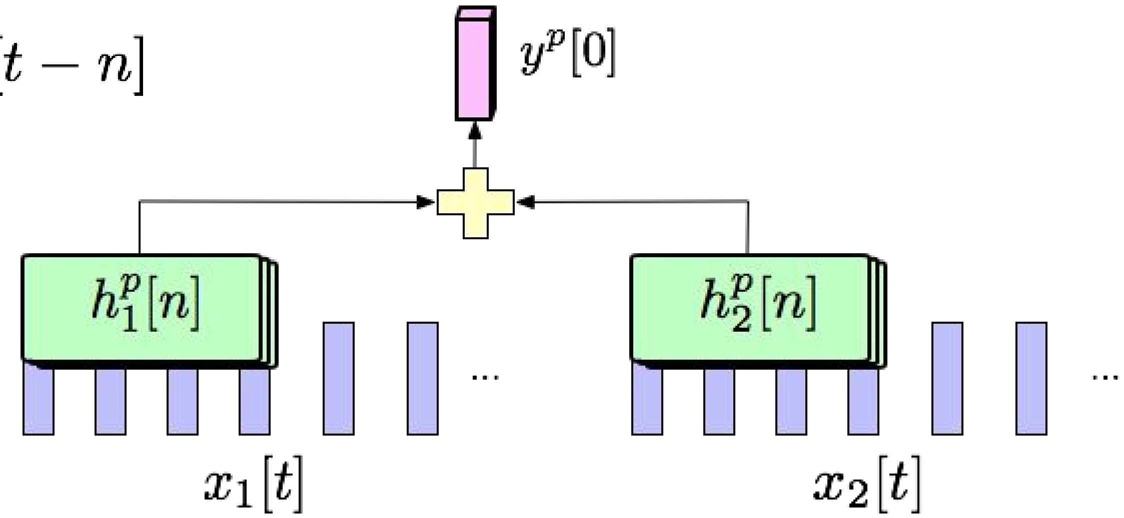
$$y[t] = \sum_{c=0}^{C} \sum_{n=0}^{N} h_c[n] x_c[t - n - \tau_c]$$

- Can we use a network to jointly estimate steering delays and filter parameters while optimizing acoustic model performance?
- *P* filters to capture many **fixed** steering delays

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

Unfactored raw-waveform architecture

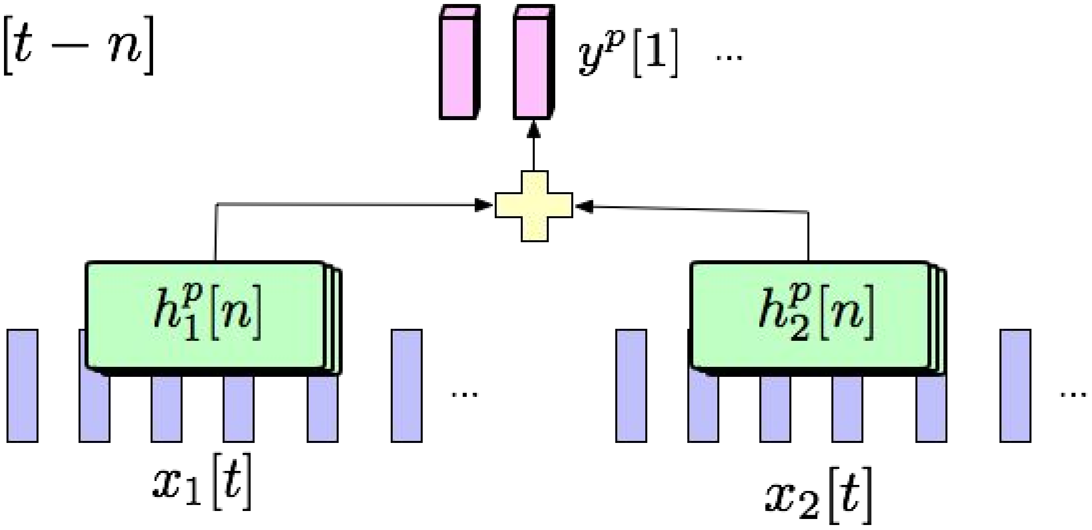
$$y^p[t] = \sum_{c=0}^{C-1} \sum_{n=0}^{N-1} h_c^p[n] x_c[t-n]$$



Layer similar to F+S but without estimating τ_c

Unfactored raw-waveform architecture

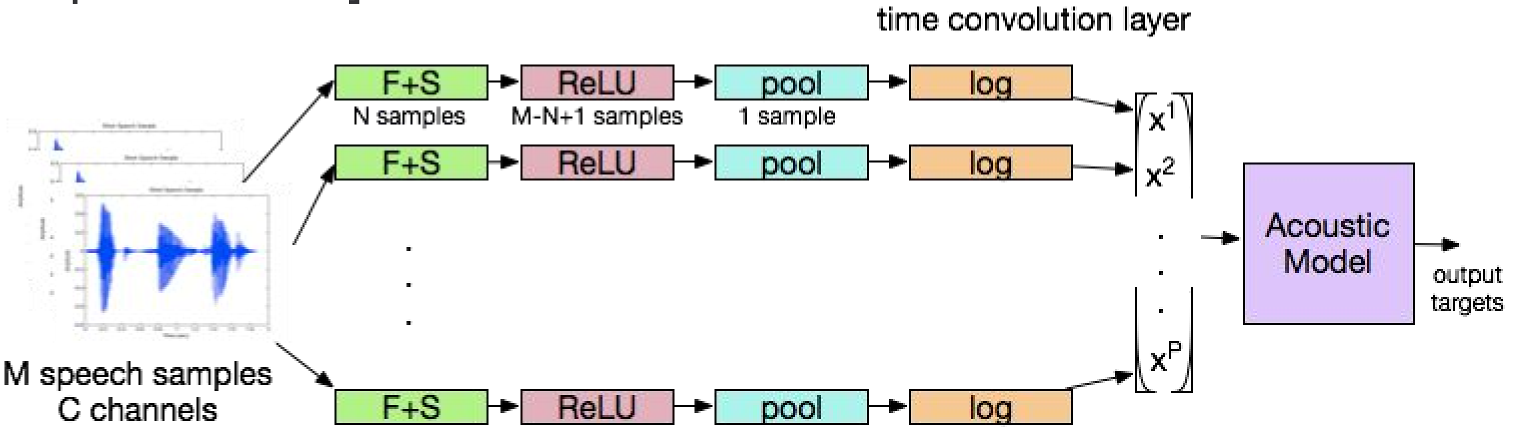
$$y^p[t] = \sum_{c=0}^{C-1} \sum_{n=0}^{N-1} h_c^p[n] x_c[t-n]$$



Layer similar to F+S but without estimating τ_c

From Samples to Time-Frequency Representation

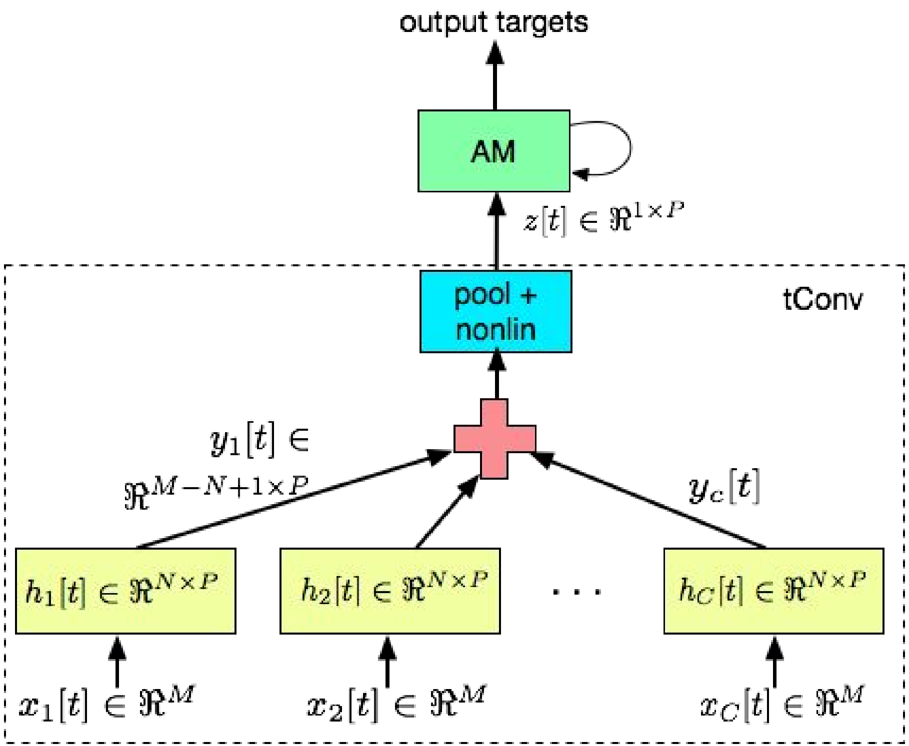
- Inspired by gammatone processing, pool the output of F+S layer to give a “time-frequency” representation invariant to short time-shifts
- 1ch raw-waveform processing explored in [T.N. Sainath et al, Interspeech 2015]



Google

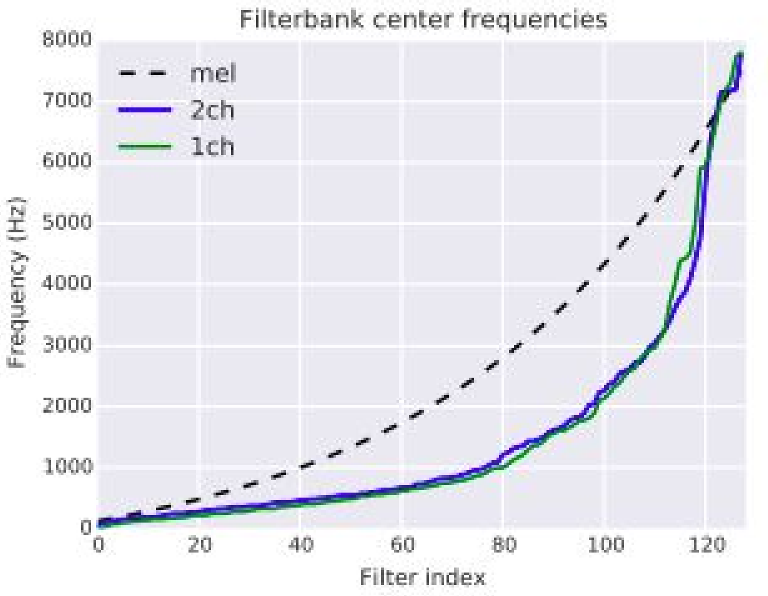
Unfactored Model

- Neural beamforming raw-waveform layer does both **spatial** **and** **spectral** filtering
- Output of this layer is passed to an AM, all layers are trained jointly!



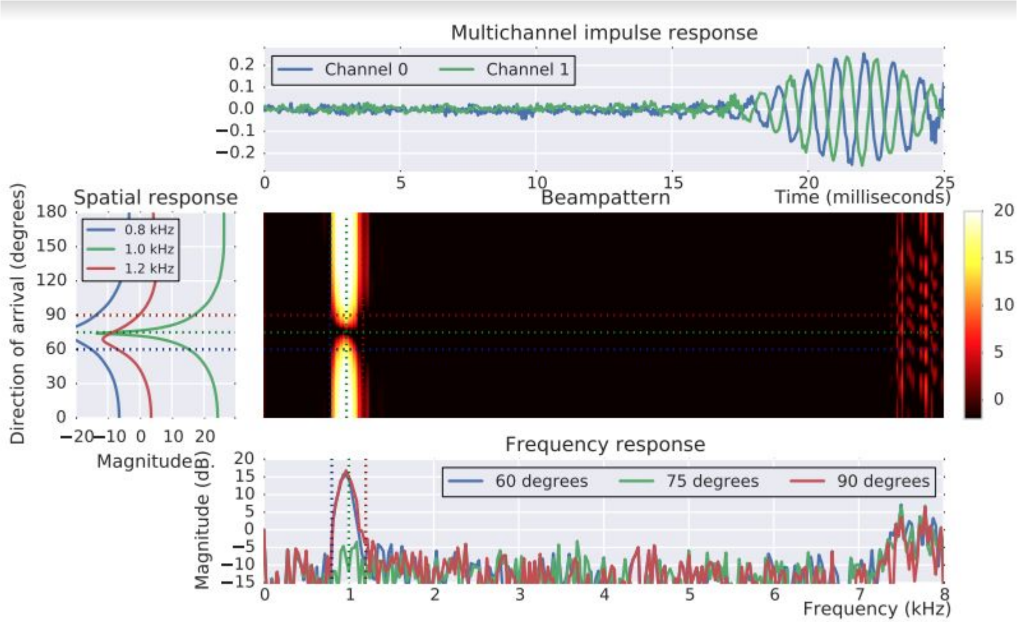
Spectral Filtering: Magnitude Response of Learned Filters

- Plot the magnitude response of the learned tConv filters
- Network seems to learn auditory-like bandpass filters
- Bandwidth increases with center frequency
- Learned filters give more resolution in lower frequencies



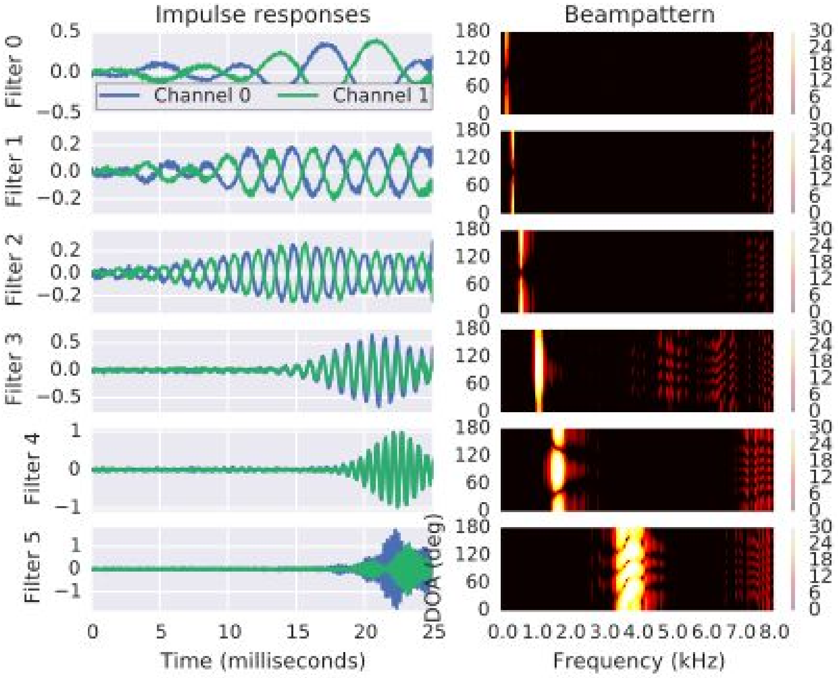
Beampattern Plots

- Pass an impulse response with different delays into filter, measure the magnitude response



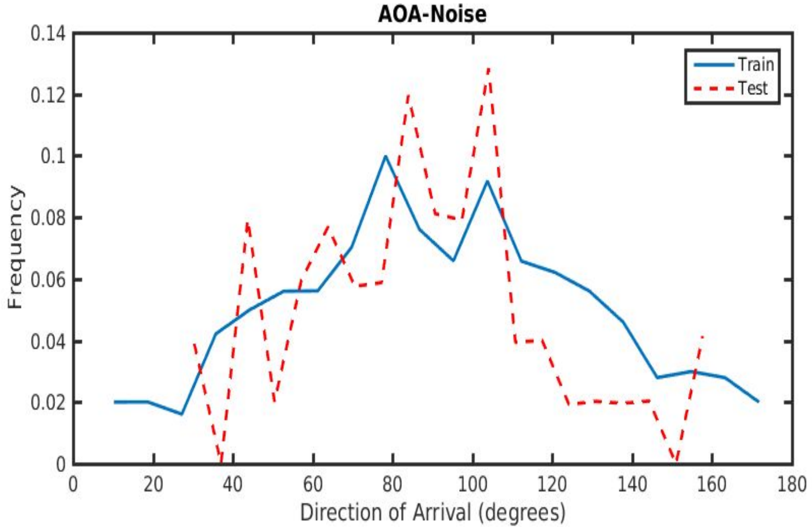
What Does The Network Learn?

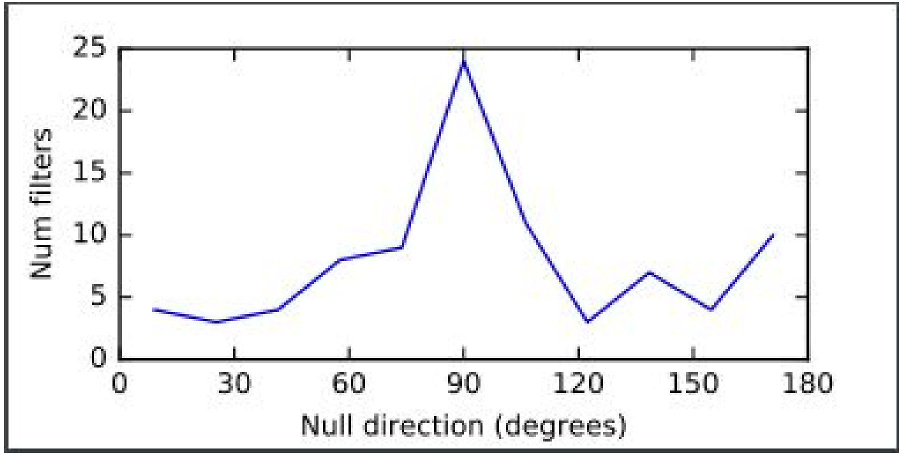
- Filter coefficients in two channels are shifted, similar to the steering delay concept.
- Most filters have bandpass response in frequency
- Filters are doing **spatial** and **spectral** filtering!



Learned Filter Null Direction

Strong correlation between AOA noise distribution and null direction of learned filters





Spatial Diversity of Learned Filters

- Increasing number of filters *P* allows more complex spatial responses
- See improvements in WER as we increase the number of spatial filters

Filters	**2ch**	**4 ch**	**8ch**
128	21.8	21.3	21.1
256	21.7	20.8	20.6
512	-	20.8	20.6

How Well Does Model Learn Localization?

- Unfactored raw-waveform, no oracle localization

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

- Delay-and-sum with oracle

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

- Time-aligned multi-channel (TAM)

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

How Well Does Model Learn Localization?

- Model trained and tested with same microphone spacing
- Unfactored raw-waveform model learns implicit localization

Feature	**1ch**	**2ch (14cm)**	**4ch (4-6-4cm)**	**8ch (2cm)**
D+S, tdoa	23.5	22.8	22.5	22.4
TAM, tdoa	23.5	21.7	21.3	21.3
raw	23.5	**21.8**	**21.3**	**21.1**

Summary, Unfactored Raw-Waveform Model

- Numbers reported after cross-entropy and sequence training
- Oracle: true target speech TDOA and noise covariance known
- Unfactored 2-channel model improves over signal channel and traditional signal processing techniques

Architecture	WER (after Seq.)
raw, 1ch	19.2
D+S, 8 channel, oracle	18.8
MVDR, 8 channel, oracle	18.7
raw, 2ch, unfactored	18.2

Factored Raw-Waveform Model

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, March 2016.

Google

Motivation

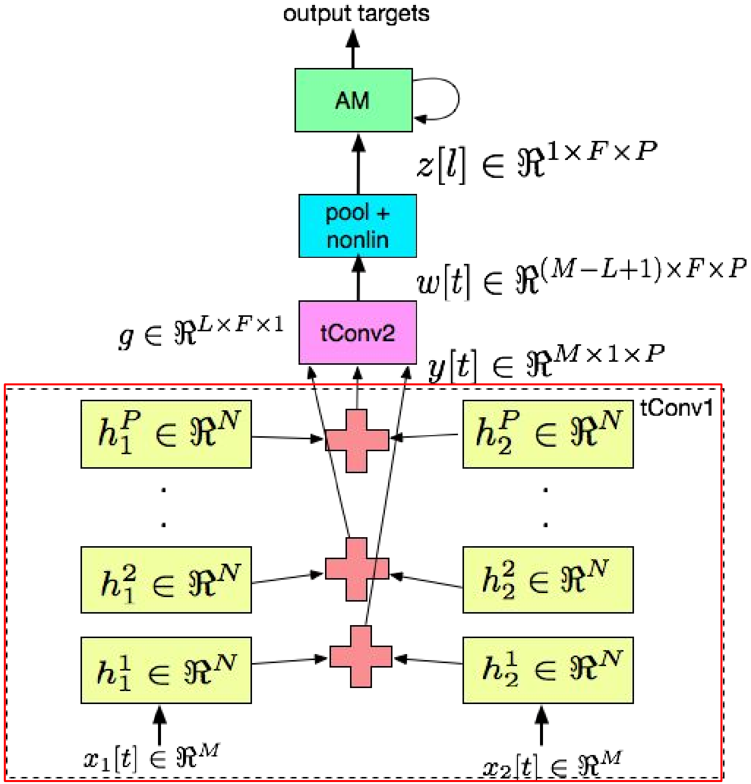
- Most multichannel systems perform **spatial filtering** separately from single channel **feature extraction**
- Unfactored raw-waveform model
  - Does spatial and spectral filtering jointly
  - Can only increase spatial directions by increasing number of filters
- Can we **factor** these operations **separately** in the network?

Spatial Layer

- We want to implement a “filter and sum” layer
- Each channel *x* is convolved with *P* short filters *h* of length *N* (i.e., 5ms)
- The outputs after convolution are combined (i.e., filter-and-sum)

$$y^p[t] = x_1[t] * h_1^p + x_2[t] * h_2^p$$

- **Factored layer** does spatial filtering in different look directions *p*

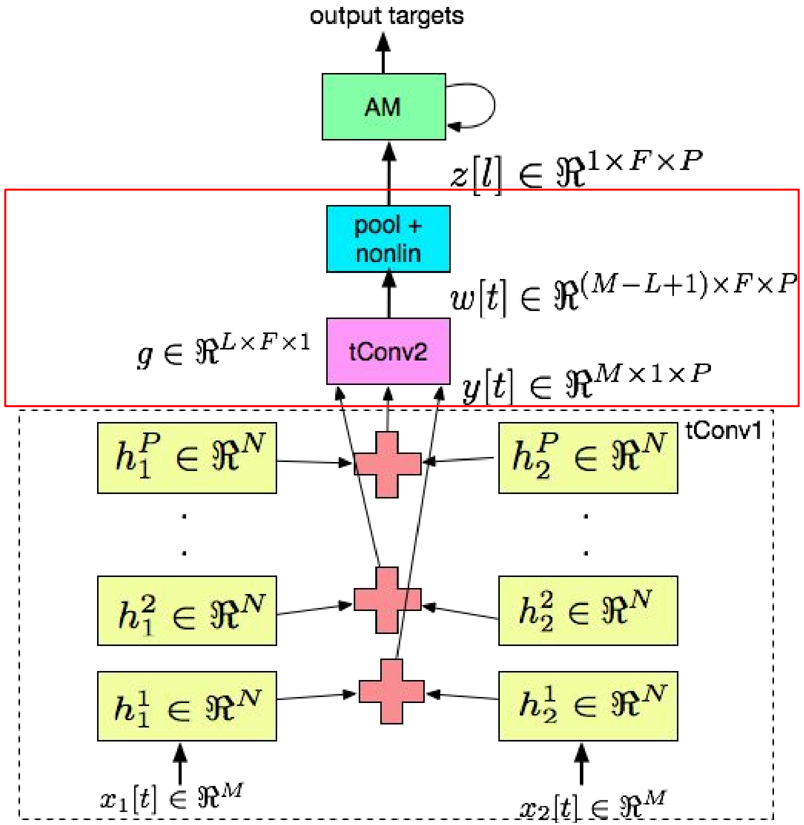


Spectral Layer

- We pass these *P* look directions to a **spectral layer** which does a time-frequency decomposition

$$w_f^p[t] = y^p[t] * g_f$$

- Factored layers are trained jointly with acoustic modeling



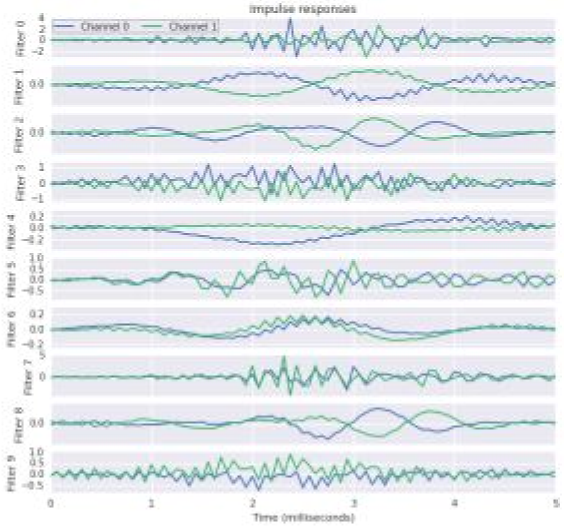
Spatial Diversity of Factored Layer

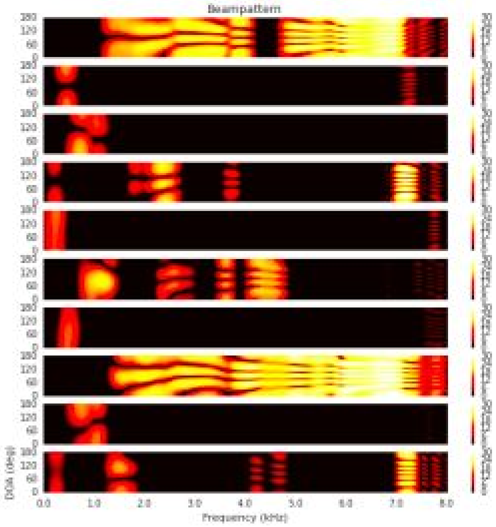
Increasing the spatial diversity of the spatial layer improves WER

# Spatial Filters *P*	WER,CE
2ch, unfactored	21.8
1	23.6
3	21.6
5	20.7
10	20.4

Spatial Analysis

- First layer is doing spatial and spectral filtering, but within broad classes!





Analysis of First Layer

- Enforce **spatial diversity** only by fixing first layer to be impulse responses at different look directions and not training the layer
- Training the layer to do spatial/spectral filtering is beneficial

First Layer	WER
Fixed (spatial only)	21.9
Trained (spatial and spectral)	20.9

Summary, Factored Raw-waveform model

- Factored network gives an additional 5% WERR over unfactored model

Architecture	**WER (after Seq.)**
raw, 1ch	19.2
D+S, 8 channel	18.8
MVDR, 8 channel	18.7
raw, 2ch, unfactored	18.2
raw, 2ch, factored	**17.2**

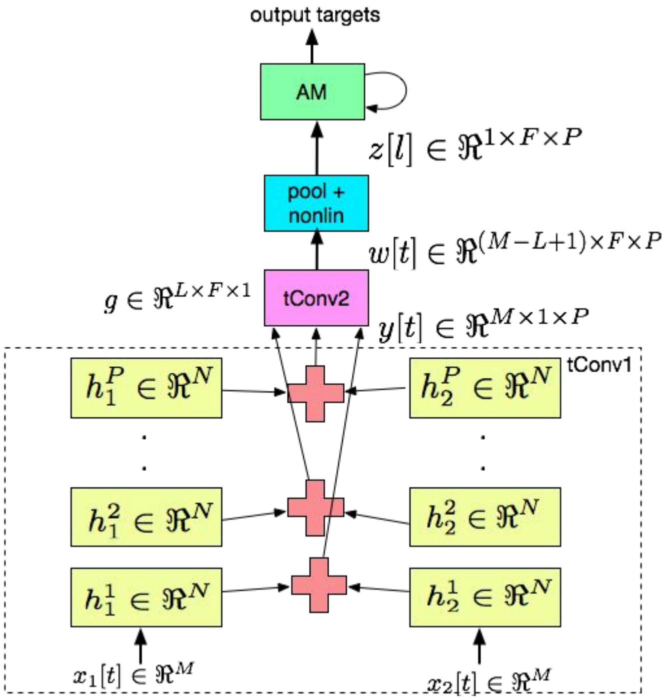
Factored CLP (fCLP) Model

T. N. Sainath, A. Narayanan, R. Weiss, E. Variani, K. Wilson, M. Bacchiani and I. Shafran, "Reducing the Computational Complexity of Multimicrophone Acoustic Models with Integrated Feature Extraction," in Proc. Interspeech, 2016.

Computational Complexity

Layer	Parameters
Input	Samples: ***M***, Channels: ***C***
Factored	Filter Size: ***N***, Look Directions: ***P***
Spectral	Filter Size: ***L***, Filters: ***F***, Filter Stride: ***S***

Layer	Total Multiplies	In Practice (*P*=5)
Spatial	$P \times C \times M \times N$	525.6K
Factored	$P \times F \times L \, x \, (M - L + 1)/S$	**62.0M**
AM	-	19.1M



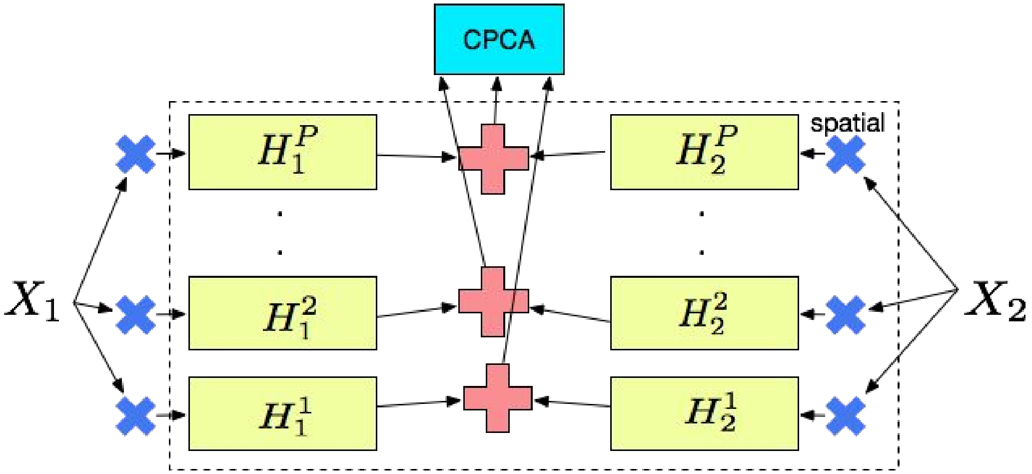
Google

Factored Model in Frequency

- Time-domain processing is expensive
- Convolution in time represented by an element-wise dot product in frequency

$$y_p[t] = \sum_{c=1}^{C} x_c[t] * h_c^p$$

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



Spectra Decomposition - Complex PCA

- Convolution in spectral layer can also be replaced by an element-wise dot product in frequency

$$w_f^p[t] = y^p[t] * g_f \longrightarrow W_f^p[l] = Y^p[l] \cdot G_f$$

- Instead of max-pooling, as is done in time, we perform **average pooling** in the frequency domain

$$Z_f^p[n] = \log \left| \sum_l Y^p[n, l] \cdot G_f[l] \right|$$

Computational Complexity Time Vs. Frequency

Layer	Parameters
Input	Samples: ***M***, Channels: ***C,*** Frequency: **K**
Factored	Filter Size: ***N***, Look Directions: ***P***
Spectral	Filter Size: ***L***, Filters: ***F***, Filter Stride: ***S***

Layer	Total Multiplies Time	Total Multiplies Frequency
Spatial	$P \times C \times M \times N$	$4 \times P \times C \times K$
Factored	$P \times F \times L \times (M - L + 1)/S$	$4 \times P \times F \times K$
AM	-	-

Google

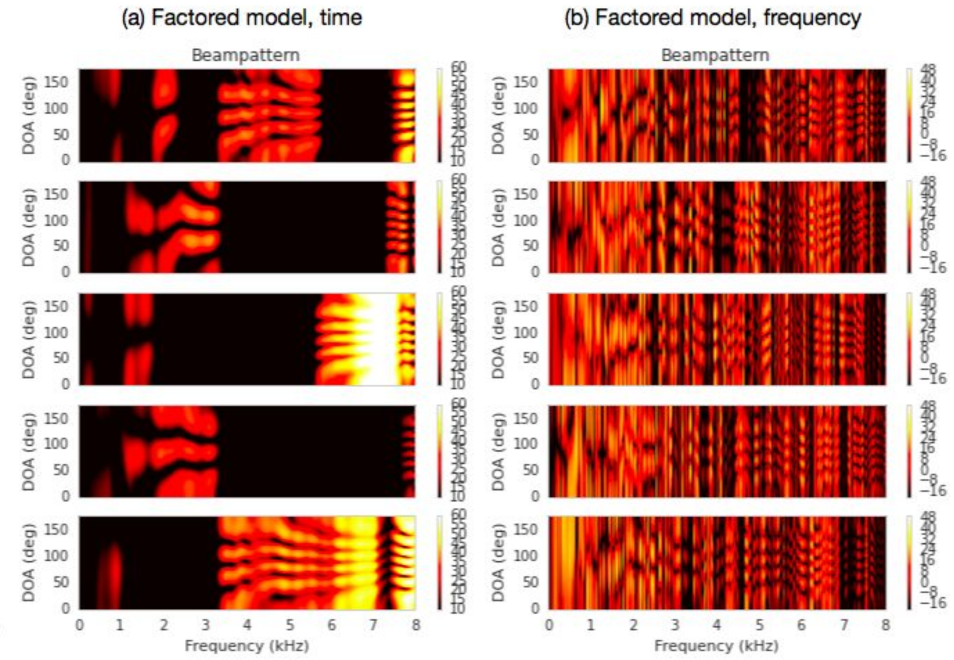
Results by Reducing Computation in Frequency

- Results with *P=5* look directions, *F=128* spectral filters
- We can reduce multiplies of the overall factored model by more than a **factor of 4** with no loss in WER

Layer	Spatial Multiplies	Spectral Multiplies	Acoustic Model	Total Multiplies	WER (Seq.)
fRaw	525.6K	62.0M	19.1M	81.6M	17.2
fCLP	**10.3K**	**655.4K**	**19.1M**	**19.7M**	**17.2**

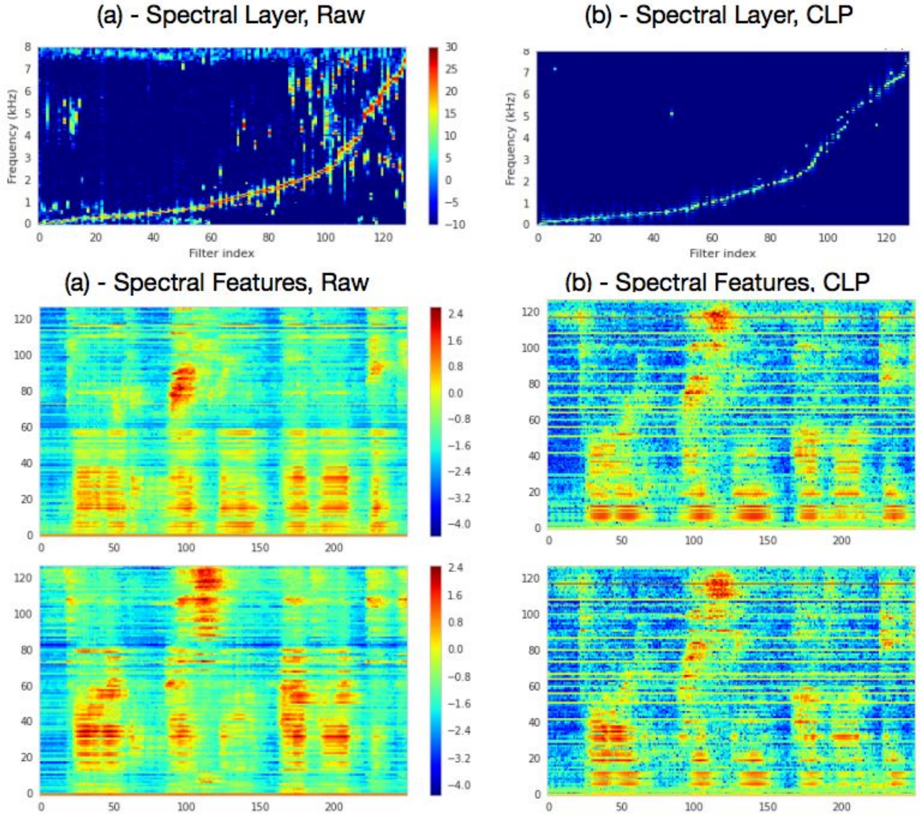
Analysis of Factored Layer

- Beampattern in time is more spatially selective than frequency



Analysis of Spectral Layer

- Magnitude response of CLP and raw-waveform are bandpass filters
- Because time modeling has more spatial selectivity at factored layer, spectral layer outputs in time more diverse compared to CLP.



Summary, fCLP

- fCLP gives improvement in computation without loss in accuracy

Architecture	WER (after Seq.)
raw, 1ch	19.2
D+S, 8 channel	18.8
MVDR, 8 channel	18.7
uRaw, 2ch	18.2
fRaw, 2ch	**17.2**
fCLP, 2ch	**17.2**

Neural Adaptive Beamforming (NAB)

B. Li, T. N Sainath, R. Weiss, K. Wilson and M. Bacchiani, "Neural Network Adaptive Beamforming for Robust Multichannel Speech Recognition," in Proc. Interspeech, 2016.

Motivation

- Thus-far all filter parameters are optimized on training data only
- It would be helpful to adapt parameters **per utterance**:
  - **Cross session variations**: Train and test mismatches cannot be reflected in those filters, such as room impulse responses different from training.
  - **Within session variations**: Dynamic changes within a single utterance cannot be address, such as moving speakers etc.
- Can we utilize statistics per training/test utterance to do adaptive beamforming similar to [Xiao et al, 2016]?

Neural Adaptive Beamforming (NAB)

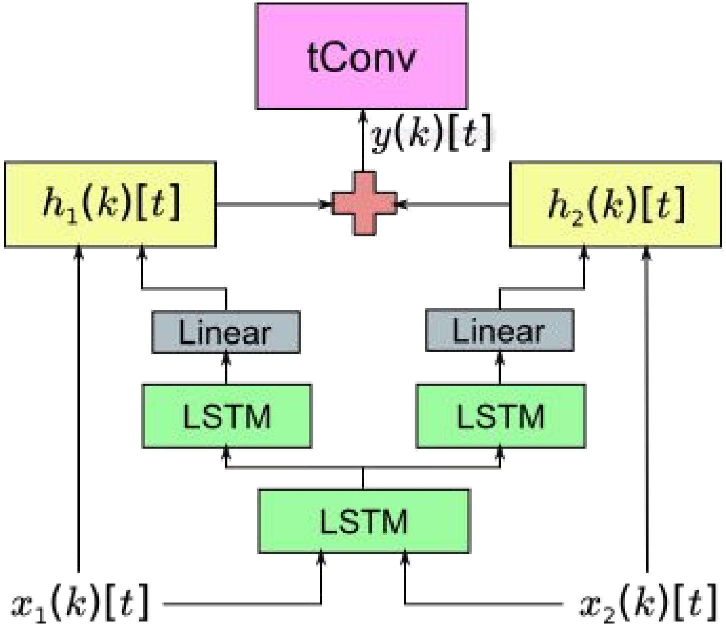
- LSTM for each channel predicts a set of filter coefficients

$$h_1(k)[t], h_2(k)[t]$$

- Convolve each channel with the filter coefficients

$$y(k)[t] = x_1(k) * h_1(k)[t] + x_2(k) * h_2(k)[t]$$

- This layer is mimicking F+S



Neural Adaptive Beamforming (NAB)

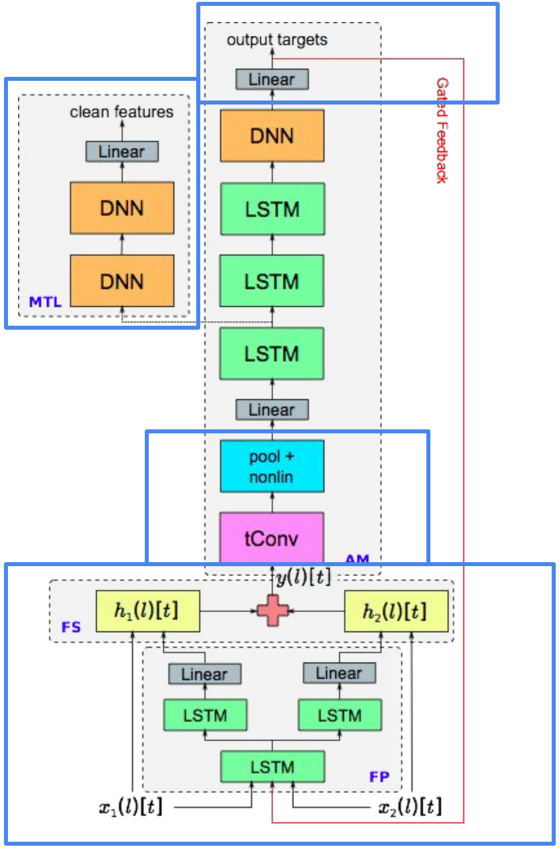
- LSTM-based adaptive beamforming
- Passed to a spectral layer to get frame-level features
- Gated history feedback

$$g^{\mathrm{fb}}(t) = \sigma(\boldsymbol{w}_x^T \cdot \boldsymbol{x}_t + \boldsymbol{w}_s^T \cdot \boldsymbol{s}_{t-1} + \boldsymbol{w}_v^T \cdot \boldsymbol{v}_{t-1})$$

Current inputs · Previous state · AM feedback

$$\left[\boldsymbol{x}_t^T, \quad g^{\mathrm{fb}}(t)\boldsymbol{v}_{t-1}^T\right]^T$$

- Denoising MTL



NAB Analysis

- Output of NAB at every frame gives a freq *x* direction *x* time beampattern
- Plot the beampattern of the NAB filters in the direction of the target speech and noise directions
- Responses in the target speech direction have relatively more speech-dependent variations than those in the noise direction

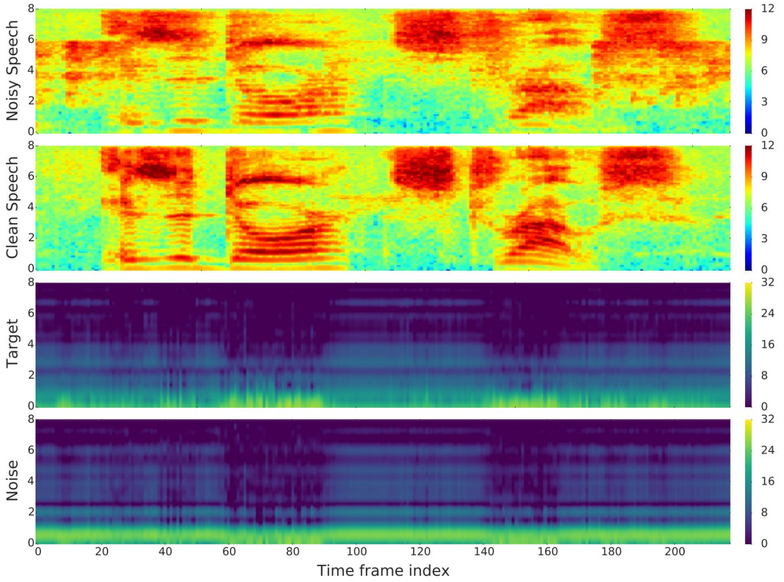


Figure 2: *Visualizations of the predicted beamformer responses at different frequency (Y-axis) across time (X-axis) at the target speech direction (3rd) and interfering noise direction (4th) with the noisy (1st) and clean (2nd) speech spectrograms.*

Google

NAB Results

- We experimented NAB in both time and frequency domain:
  - NAB in time matches factored model
  - NAB in frequency degrades as too many filter coefficients to estimate

Method	CE WER
fRaw, time	20.4
NAB, time	**20.5**
fCLP, freq	20.5
NAB, freq	21.0

Summary, NAB Model

- NAB model matches performance of factored models

Architecture	WER (after Seq.)
raw, 1ch	19.2
D+S, 8 channel	18.8
MVDR, 8 channel	18.7
uRaw, 2ch	18.2
fRaw, 2ch	**17.2**
fCLP, 2ch	**17.2**
NAB, 2ch	**17.2**

Results on More Realistic Data

T. N. Sainath, R. J. Weiss, K. W. Wilson, B. Li, A. Narayanan, et al, "Multichannel Signal Processing with Deep Neural Networks for Automatic Speech Recognition," in IEEE Transactions on Speech and Language Processing, 2017.
B. Li, T. N. Sainath, J. Caroselli, A. Narayanan, M. Bacchiani, et al, "Acoustic Modeling for Google Home," in Proc. Interspeech, 2017.

Experimental Setup, re-recorded Data

Training data:

- 22M English utterances
- 18,000 hours noisy data
- artificially corrupted with music, ambient noise, recordings of "daily life" environments
- SNRs: 0 ~ 30dB, avg. = 11dB
- Reverberation RT60: 0 ~ 900ms, avg. = 500ms
- 2 channel microphone distance: 71mm

Testing data:

- 13K English utterances
- 15 hours data
- rerecorded:
  - SNRs: 0 ~ 20dB
  - RT60: ~200ms
  - Rev-I: mic on coffee table
  - Rev-II: mic on TV stand
- 2 channel microphone distance: 75mm

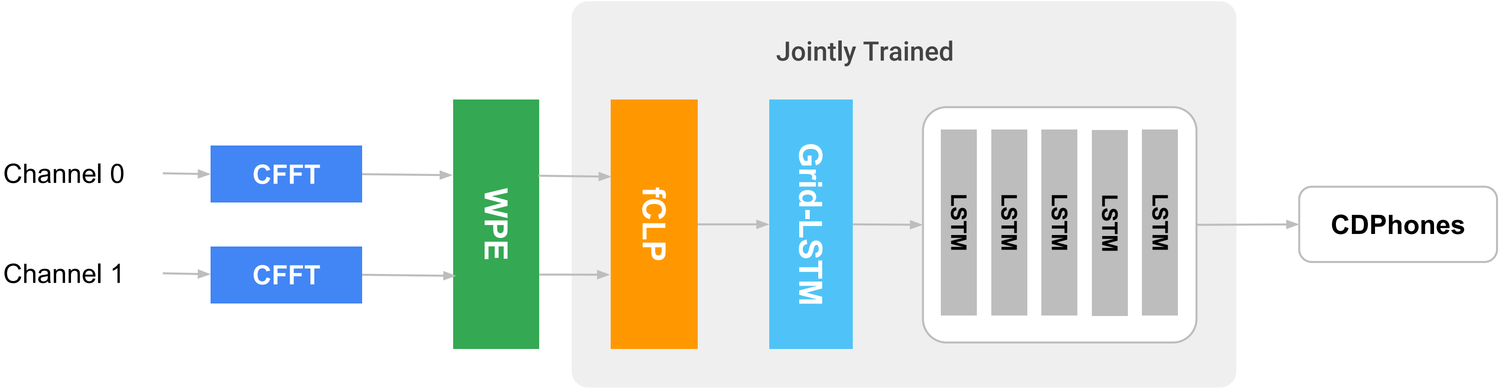
Re-recorded Results

- On rerecorded sets, can get a 10-14% relative improvement with 2 channel **fRaw**, **fCLP** over single channel
- 2ch **fRaw**, **fCLP** matches the performance of a **7 ch oracle superdirective beamformer**
- Google HOME is designed with **2 microphones** to do server-side recognition

Method	Rev I	Rev II	Rev I Noisy	Rev II Noisy	Ave
raw, 1ch	18.6	18.5	26.7	26.7	22.9
uRaw, 2ch	17.9	25.9	24.7	24.7	21.5
fRaw, 2ch	**17.1**	**24.6**	**24.2**	**24.2**	**20.7**
fCLP, 2ch	**17.4**	**25.2**	**23.5**	**23.5**	**20.7**
NAB, 2ch	17.8	18.1	27.1	26.1	22.3
7 ch, oracle superdirective	-	-	25.3	23.7	

Google HOME System Overview

- Take what we learned on simulated and re-recorded data and apply to Google HOME data [Li, IS-2017]
- Input is CFFT features for time efficiency
- Weighted Prediction Error (WPE) to reduce reverberation [Caroselli, IS-2017]
- Neural beamforming uses fCLP, which gave best tradeoff between computation and WER
- Grid-LSTM to model time-frequency correlations [Sainath, IS-2016; Li, IS-2017]



WER on Google HOME Traffic

- Setup:
  - Model trained on 22,000 simulated noisy VS utterances
  - The final system: **WPE** + **fCLP** + **Grid-LSTM**
  - Cross-Entropy + Sequence training
  - Google Home real test set, representative of real traffic
- A **16% overall** WER reduction on live Google HOME data
- Major win comes in noisy environments:
  - **26%** WER reduction in **speech background** noise
  - **18%** WER reduction in **music** noise

Model	full	clean	Noise Type		
			speech	music	Other
Baseline (log-mel)	6.1	5.1	8.5	6.2	6.0
Proposed	5.1	4.9	6.3	5.1	5.0
rel.	-16.4	-3.9	-25.9	-17.7	-16.7

Table 4. WERs for the proposed Google Home system(with sequence training).

In-Domain Tuning

- Continue sequence training on **4,000 hours** in-domain data
- **Another 4%** relative improvements
- **Overall, a 8~28%** relative improvement over the baseline system.
- WER of Google HOME is around **4.9%** on live data!

Model	full	clean	Noise Type		
			speech	music	Other
Baseline (log-mel)	6.1	5.1	8.5	6.2	6.0
Proposed	5.1	4.9	6.3	5.1	5.0
Proposed + Adaptation	**4.9**	**4.7**	**6.1**	**4.9**	**4.8**
rel.	**-3.9**	**-4.1**	**-3.2**	**-3.9**	**-4.0**

Table 5. WERs for the proposed Google Home system with adaptation.

Future Directions

- Google HOME works relatively well but there are areas to improve
- Multi-talker scenarios
- Using multiple modalities to improve robustness
- Multi-channel in end-to-end framework (similar to [Ochiai 2017])

Conclusions

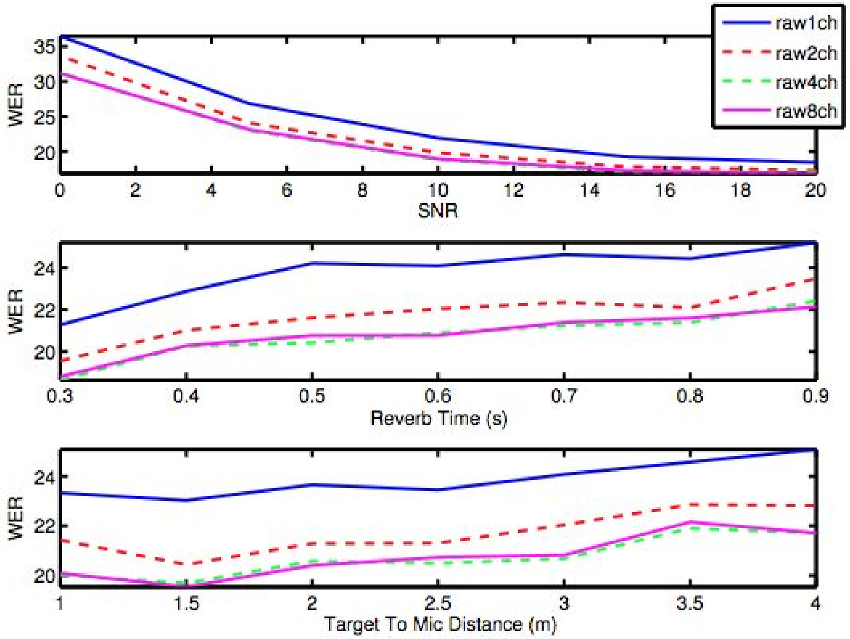
- Overview of Various Multichannel Architectures
- Neural beamforming architectures include
  - Unfactored raw-waveform - uRaw
  - Factored raw-waveform - fRaw
  - Factored Complex Linear Prediction - fCLP
  - Neural Adaptive Beamforming - NAB
- fCLP achieves best tradeoff between WER and time and is used in Google HOME

References

- T. N. Sainath, R. J. Weiss, A. Senior, K. W. Wilson 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 Location and Microphone Spacing Invariant Acoustic Modeling from Raw Multichannel Waveforms," in Proc. ASRU, December 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, March 2016.
- T. N. Sainath, A. Narayanan, R. Weiss, E. Variani, K. Wilson, M. Bacchiani and I. Shafran, "Reducing the Computational Complexity of Multimicrophone Acoustic Models with Integrated Feature Extraction," in Proc. Interspeech, 2016.
- B. Li, T. N Sainath, R. Weiss, K. Wilson and M. Bacchiani, "Neural Network Adaptive Beamforming for Robust Multichannel Speech Recognition," in Proc. Interspeech, 2016.
- E. Variani, T. N. Sainath, I. Shafran and M. Bacchiani, "Complex Linear Projection (CLP): A Discriminative Approach to Joint Feature Extraction and Acoustic Modeling," 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.
- 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 "Raw Multichannel Processing Using Deep Neural Networks," chapter in New Era for Robust Speech Recognitino: Exploiting Deep Learning, 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.

Backup

Multi-channel WER Breakdown



Multi-microphone processing helps to enhance signal and suppress noise