Raw Waveform Modelling for ASR

A Literature Review

Part III

Erfan Loweimi

Centre for Speech Technology Research (CSTR)
The University of Edinburgh

Listen! 12.2.2020

Raw Waveform Acoustic Modelling

- Divide-and-conquer paradigm may not be needed ...
  - Solve feature extraction & AM problems simultaneously
- Advantages
  - Task-specific features, employ all info, learn basis functions, mid-term processing, do not need exact alignment
- Challenges
  - Learning in High-dim feature, discard prior knowledge, ...

Part I – Summary

- Conventional features are still better
- Architecture is important (CNN rather than MLP)
- Data amount and activation function can narrow the gap
- Interpretability
  - First layer → time-frequency analysis
  - Second layer → modulation spectrum processing
  - Filters resemble auditory filters
    - More filters in low freq, wider filters in high frequencies (trend-wise)

Part II – Summary

- Baidu → Multi-resolution CNNs
- JHU → NIN + iVector + Normalisation + Data augment
- Cambridge → Multi-Span CNNs
- Google → CLDNN Acoustic Modelling
  - CNN + LSTM + MLP
  - Goal: super-additive combination

Our Plan ...

- Part **I** → IDIAP + AACHEN
- Part **II** → Multi-Resolution + Google
- Part **III** → Google
- Part IV → Parametric CNNs

Learning the Speech Front-end With Raw Waveform CLDNNs

Tara N. Sainath, Ron J. Weiss, Andrew Senior, Kevin W. Wilson, Oriol Vinyals

Google, Inc. New York, NY, U.S.A

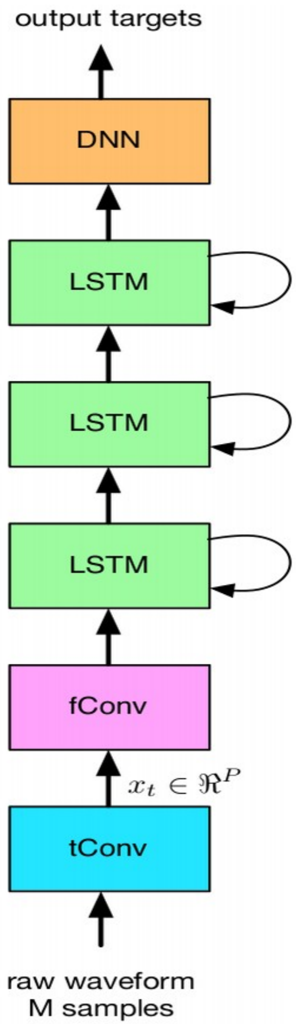
{tsainath, ronw, andrewsenior, kwwilson, vinyals}@google.com

INTERSPEECH 2015
September 6 - 10
Dresden, Germany

Raw waveform CLDNN

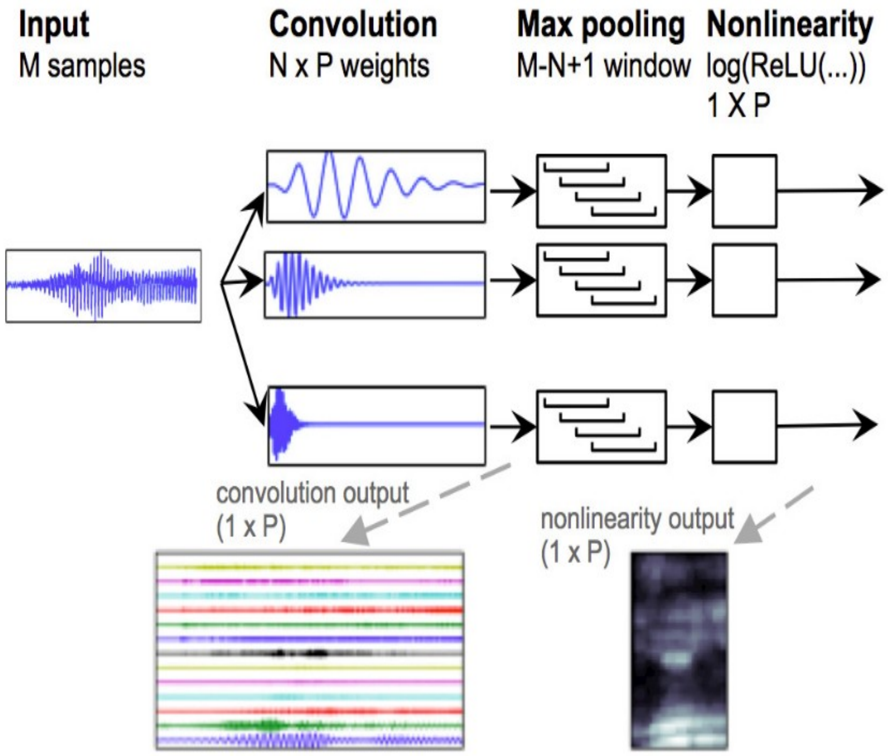
- tConv → conv in time
- fConv → conv in freq
- LSTM: dynamic modelling
- DNN: abstraction
  - 1 FC layer with 1024 units

- Output of tConv, x_t, is passed to fConv without temporal context → "*... not to help on larger data sets.*"



Convolution in time → tConv

- tConv layer consists of
  - bank of bandpass FIR filters
  - pooling + non-linearity
- Feature maps ≡ “frequency”
- Feature size
  - 1 x M → M-N+1 x P → 1 x P
- Non-linearity: log(ReLU(.) + 0.01)
- Output ≡ CRBE



M: input size
P: #filters (40)
N: Filter length

Experimental Results

- Frame length, M: 25 → 35ms
  - 3.5 abs WER reduction
- Gammatone (GT) init.
  - $WER_{GT} = WER_{random} - 0.2$
  - If frozen => $WER_{GT} = WER_{random}$
- Max-Pooling → lower WER
  - "… *MaxP emphasises transients ... p-norm and AveP smooth out* ..."

N is fixed in 25ms

Filter Size (N (ms))	Window Size (M (ms))	Init	WER
400 (25ms)	400 (25ms)	random	19.9
400 (25ms)	560 (35ms)	random	16.4
400 (25ms)	560 (35ms)	gammatone	**16.2**
400 (25ms)	560 (35ms)	gammatone untrained	16.4

Method	WER
max	16.2
l_2	16.4
average	16.8

data: ~ 2000 h MTR

Raw vs Log-Mel Features

- Equal performance for
  - $C_1L_3D_1$, $C_1L_2D_1$, L_3D_1
  - Best WER → $\mathbf{C_1L_3D_1}$
- Fbank better for $C_1L_1D_1$ & D_6
  - Dynamics modelling Capacity(?)

$C_xL_yD_z$ → x,y,z #layers

Feature	Model	WER
log-mel	C1L3D1	16.2
raw	C1L3D1	16.2
log-mel	L3D1	16.5
raw	L3D1	16.5
raw	L3D1, rand init	16.5
log-mel	C1L2D1	16.6
raw	C1L2D1	16.6
log-mel	C1L1D1	17.3
raw	C1L1D1	17.8
log-mel	D6	22.3
raw	D6	23.2

– data: ~ 2000 h MTR
– Raw always has tConv

Raw vs Log-Mel Features

- Clean vs Noisy
  - test matched
  - Clean → Slightly better
- raw+log-mel is super-additive
- Effect of data amount
  - No clear trend! Similar WER
- Seq training is better than CE

data: ~ 2000 h

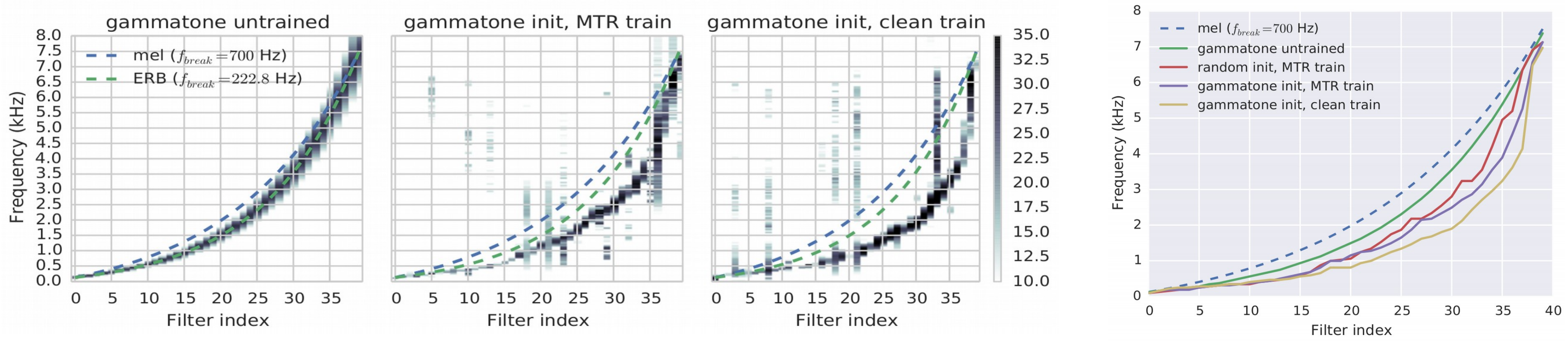
Training Set	Feature	WER - CE	WER - Seq
Clean	log-mel	14.0	12.8
Clean	raw	13.7	12.7
MTR	log-mel	16.2	14.2
MTR	raw	16.2	14.2

data: ~ 2000 h

Feature	WER - CE	WER - Seq
raw	16.2	14.2
log-mel	16.2	14.2
raw+log-mel	**15.7**	**13.8**

Hrs	WER-raw	WER-log-mel
666	18.8	18.4
1,333	17.1	17.3
2,000	16.2	16.2
40,000	15.5	15.4

Filter Interpretation



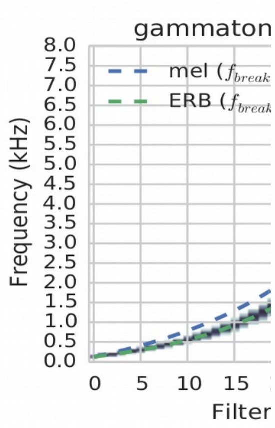
– Similar to auditory filters

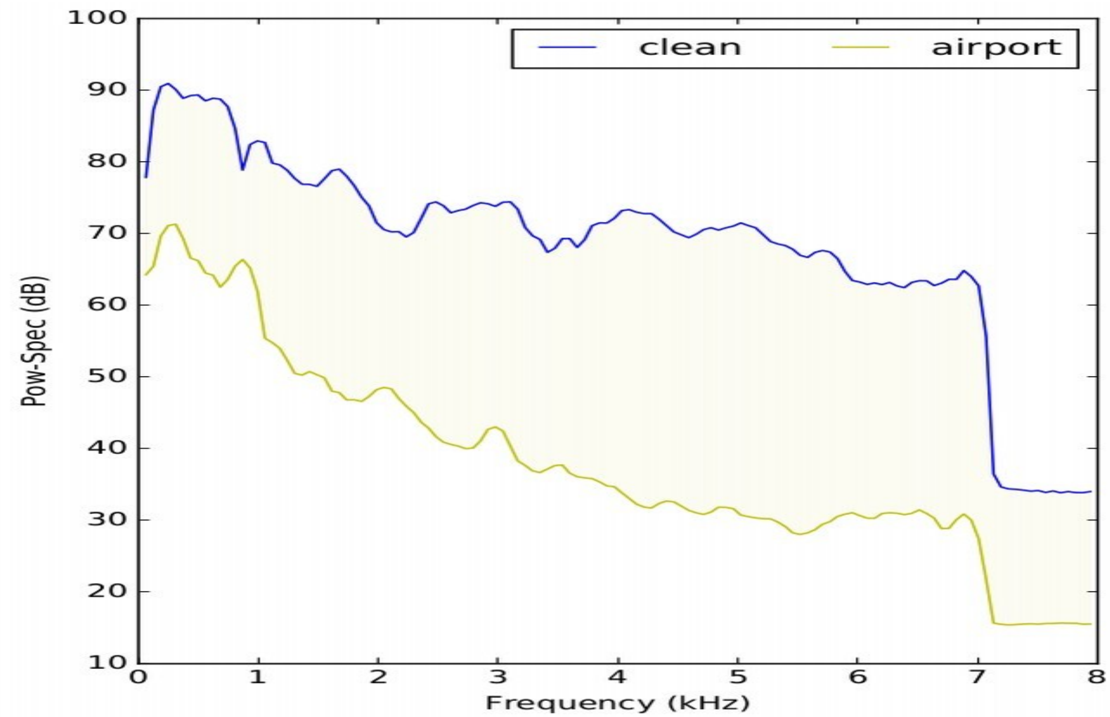
* BW increases with f_c (trend-wise)
* More filters in low freq → higher resolution and selectivity

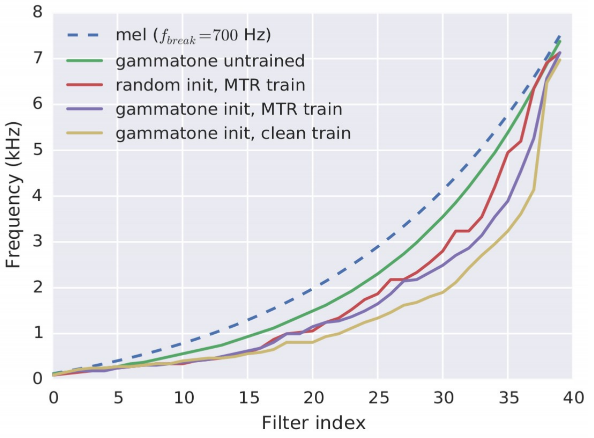
– Clean vs Noisy

* For noisy more filters in high frequencies

Filter Interpretation







ectivity

– Clean vs Noisy

  * For noisy more filters in high frequencies

SPEECH ACOUSTIC MODELING FROM RAW MULTICHANNEL WAVEFORMS

Yedid Hoshen[1], *Ron J. Weiss*[2], *and Kevin W. Wilson*[2]

[1]Hebrew University of Jerusalem, Jerusalem, Israel
[2]Google Inc, New York, NY, USA
ydidh@cs.huji.ac.il, {ronw,kwwilson}@google.com

SPEAKER LOCATION AND MICROPHONE SPACING INVARIANT ACOUSTIC MODELING FROM RAW MULTICHANNEL WAVEFORMS

Tara N. Sainath, Ron J. Weiss, Kevin W. Wilson, Arun Narayanan, Michiel Bacchiani and Andrew Senior

Google, Inc., New York, NY, USA
{tsainath, ronw, kwwilson, arunnt, michiel, andrewsenior}@google.com

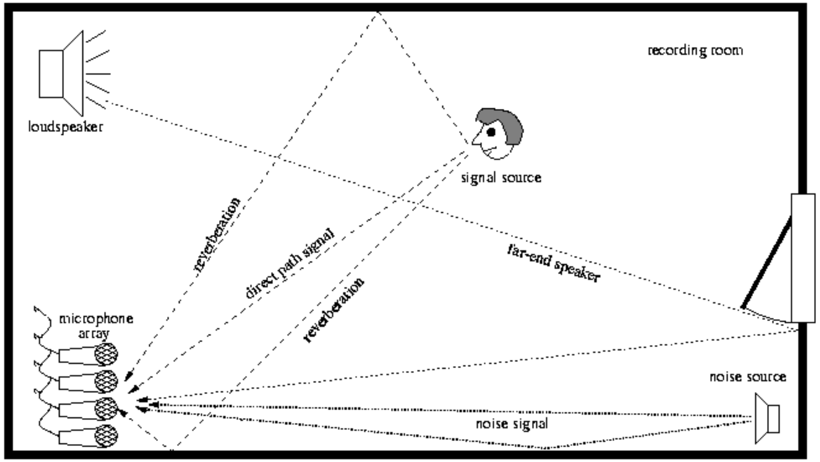
FACTORED SPATIAL AND SPECTRAL MULTICHANNEL RAW WAVEFORM CLDNNS

Tara N. Sainath, Ron J. Weiss, Kevin W. Wilson, Arun Narayanan, Michiel Bacchiani

Google, Inc., New York, NY, USA
{tsainath, ronw, kwwilson, arunnt, michiel}@google.com

E. Loweimi

A Brief Review of Beamforming



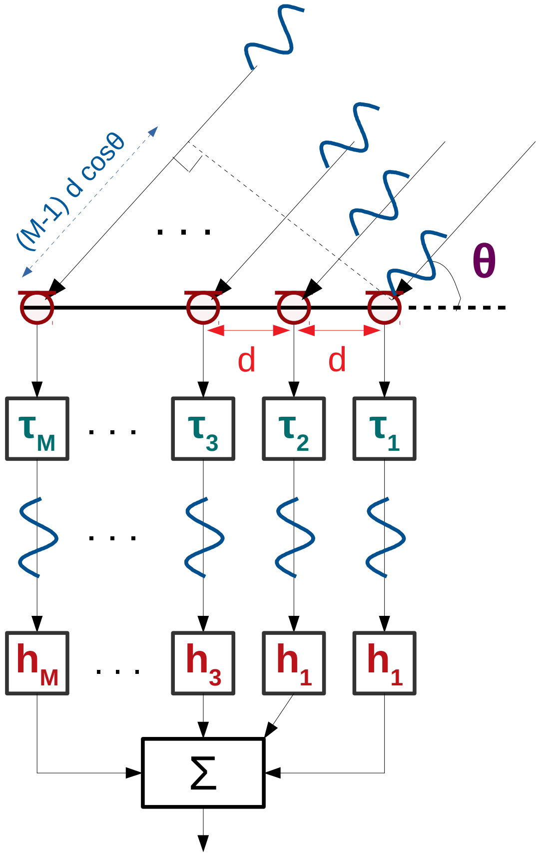
E. Loweimi

Beamforming ≡ Spatial Filtering

- Delay-and-Sum (DaS)
  - Delay ≡ align (synchronise)
  - $\tau \leftarrow \theta \leftarrow$ Localisation
- Filter-and-Sum (FaS)
  - Y? beam pattern shaping, etc.

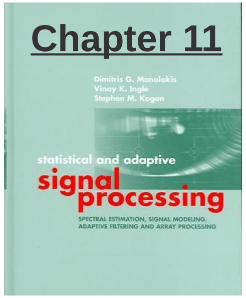
$$y[t] = \sum_{m=1}^{M} h_m[t] * x_m[t - \tau_m]$$

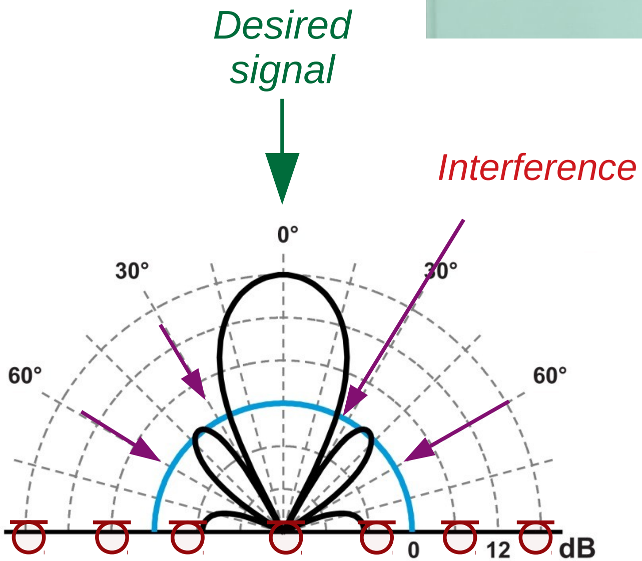
$$\tau_m = (m-1)\frac{d\cos\theta}{\text{v}}$$



Some BeamForming Jargon ...

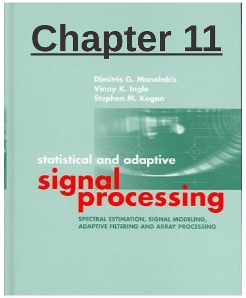
- TDoA, DOA, Steering vector, **Null**
- Broadside, Aperture, Azimuth, Elevation
- Spatial freq, Nyquist sampling, Resolution
- Uniform Linear Array (ULA)
- SINR, MVDR
- narrowband assumption
- Far-field, Near-field
- BF domain; time or frequency?





MVDR Beamforming

- **Minimum Variance Distortionless Response**
- Goal: minimise SINR
- IDEAL solution requires ...
  - Desired signal direction, θ_s
  - Interference and noise corr mat, R_{i+n}
- PRACTICAL: Recursive + Est R_{i+n}
  - Training data
  - Diagonal loading



Chapter 11

$$\mathbf{x}(t) = \mathbf{s}(t) + \mathbf{i}(t) + \mathbf{n}(t)$$
$$\mathbf{s}(t) = s(t)\mathbf{a}(\theta_s)$$
$$\mathbf{a}(\theta_s) = [1, e^{-j\omega_c \tau_2}, ..., e^{-j\omega_c \tau_M}]^T$$
$$\mathbf{y}(t) = \mathbf{w}^H \mathbf{x}(t)$$
$$SINR = \frac{\mathbb{E}\{|\mathbf{w}^H \mathbf{s}|^2\}}{\mathbb{E}\{|\mathbf{w}^H(\mathbf{i}+\mathbf{n})|^2\}} = \frac{\sigma_s^2 |\mathbf{w}^H \mathbf{a}(\theta_s)|^2}{\mathbf{w}^H \mathbf{R}_{i+n} \mathbf{w}}$$

$$\mathbf{w}^* = \text{argmin } \mathbf{w}^H \mathbf{R}_{i+n} \mathbf{w}$$
$$\text{s.t. } \mathbf{w}^H \mathbf{a}(\theta_s) = 1$$

$$\mathbf{w}^* = \alpha \, \mathbf{R}_{n+i}^{-1} \, \mathbf{a}(\theta_s)$$

BeamForming for Far-field ASR

- Classic approach → Signal Processing
  - Localisation + Beamforming + post-filtering + Acoustic Modelling
  - Done independently

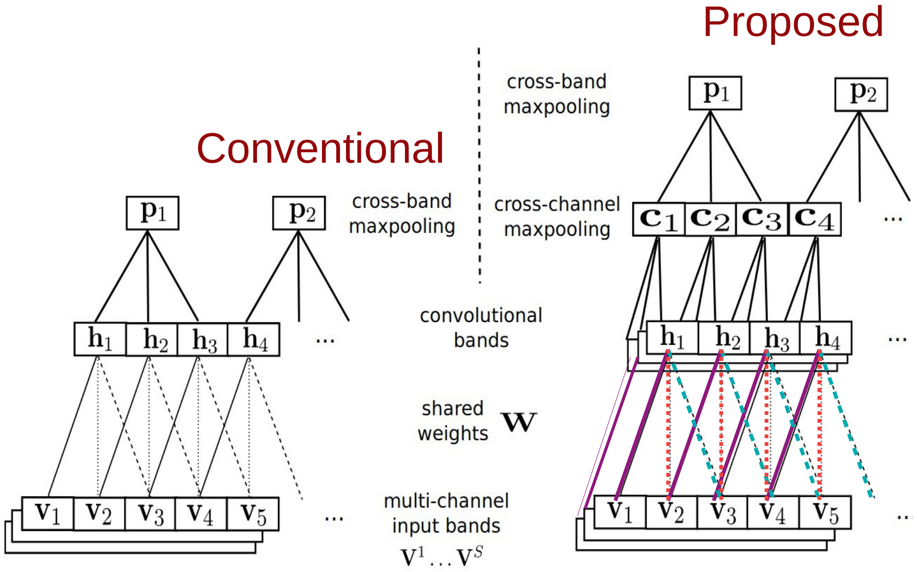
- Modern approach → Learning + DNN
  - GMM-HMM framework
    - Seltzer et al, LIMABEAM (LIkelihood-Maximising BEAMforming), 2004
  - Neuro BF
    - Swietojanski et al, CNNs for DSR, 2014
    - Hoshen et al, Google, 2015
    - Tara Sainath et al, Google, 2015→ 2017

1120 IEEE SIGNAL PROCESSING LETTERS, VOL. 21, NO. 9, SEPTEMBER 2014

Convolutional Neural Networks for Distant Speech Recognition

Pawel Swietojanski, *Student Member, IEEE*, Arnab Ghoshal, *Member, IEEE*, and Steve Renals, *Fellow, IEEE*

- Using CNN for BF & AM*
- BeamForming
  - 1) Channel-wise Conv
    - filters tied across channels
  - 2) Two-way pooling
    - 2.1) Cross-channel pooling
    - 2.2) cross-band pooling



– DNN → Fbank, 6 H Layers, 2048, Sigmoid
– CNN → Conv + 5L-DNN, J=128, F=9, L=1

SPEECH ACOUSTIC MODELING FROM RAW MULTICHANNEL WAVEFORMS

Yedid Hoshen[1], *Ron J. Weiss*[2], *and Kevin W. Wilson*[2]

[1]Hebrew University of Jerusalem, Jerusalem, Israel
[2]Google Inc, New York, NY, USA
ydidh@cs.huji.ac.il, {ronw,kwwilson}@google.com

SPEAKER LOCATION AND MICROPHONE SPACING INVARIANT ACOUSTIC MODELING FROM RAW MULTICHANNEL WAVEFORMS

Tara N. Sainath, Ron J. Weiss, Kevin W. Wilson, Arun Narayanan, Michiel Bacchiani and Andrew Senior

Google, Inc., New York, NY, USA
{tsainath, ronw, kwwilson, arunnt, michiel, andrewsenior}@google.com

FACTORED SPATIAL AND SPECTRAL MULTICHANNEL RAW WAVEFORM CLDNNS

Tara N. Sainath, Ron J. Weiss, Kevin W. Wilson, Arun Narayanan, Michiel Bacchiani

Google, Inc., New York, NY, USA
{tsainath, ronw, kwwilson, arunnt, michiel}@google.com

SPEECH ACOUSTIC MODELING FROM RAW MULTICHANNEL WAVEFORMS

Yedid Hoshen[1], *Ron J. Weiss*[2], *and Kevin W. Wilson*[2]

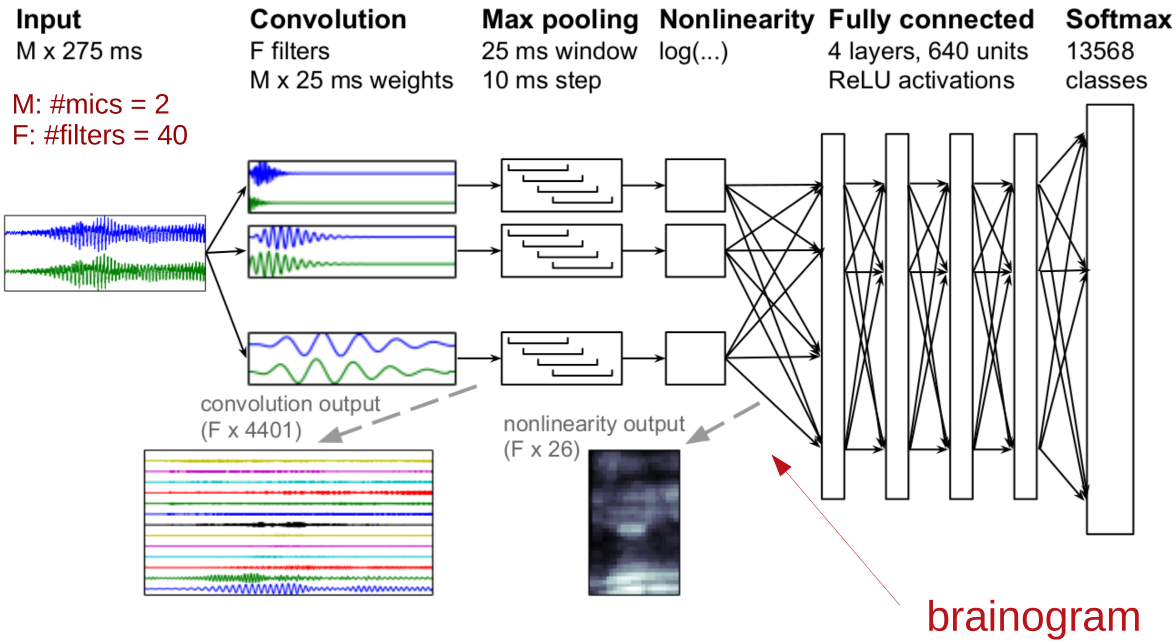
[1]Hebrew University of Jerusalem, Jerusalem, Israel
[2]Google Inc, New York, NY, USA
ydidh@cs.huji.ac.il, {ronw,kwwilson}@google.com

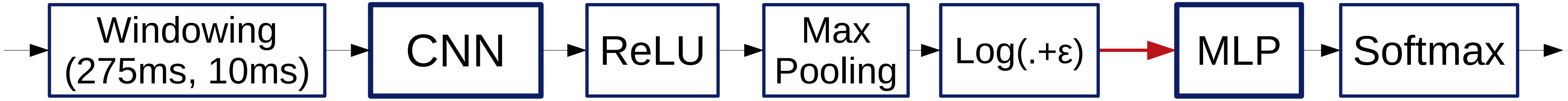
ICASSP 2015
Brisbane Australia

IDEA and Contribution

- Single-channel & Multi-channel Raw waveform AM
- Multi-channel
  - Joint AM & Localisation + Beamforming
- First Conv layer
  - Joint Spatial and Spectral filtering
- Advantage over log-mel feature
  - Phase information → better localisation & Beamforming

Architecture



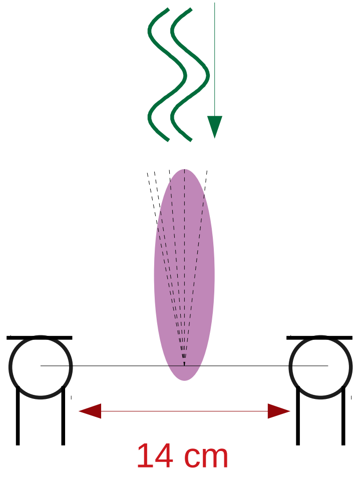


Experimental Results

- Frame Blocking, 275ms, 10 ms hopping, fbank → 40D, 26 stacked frames
- Conv layer → time-freq analysis, F x M x 25ms (F: #filters=40, M: #mics=2)
- Rectification and max-pooling (across time, separately for each filter over a window of 25 ms hopped by 10 ms
- Compressive non-linearity → log(. + 0.01)
  - 0.01 offset → Numerical stability, dynamic range compression
- MLP → Fully-connected, 4 layers with 640 ReLU units
- Softmax nodes → 13568 tied CD state unites
- Training data normalised to have zero mean and unit variance
- data: 400/36 hours for train/test, clean, Google Voice Search
- Optimisation: ASGD, Adagrad with 0.01 learning rate, batch size: 100, #epochs: 13

Setup – Training data

- Data: Voice Search + noise (YouTube) + room simulator
- Clean: Train: 400 h; testset: 36 h
- Fixed: clean + room simulator
  - additive noise (5--25dB) and reverb ($RT_{60} < 400$ms)
- Varied: speakers position is varied
  - Target speaker: Rand ±5º of broadside
  - Noise direction: Rand ±90º

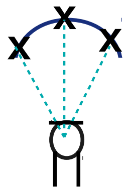


Single-Channel System

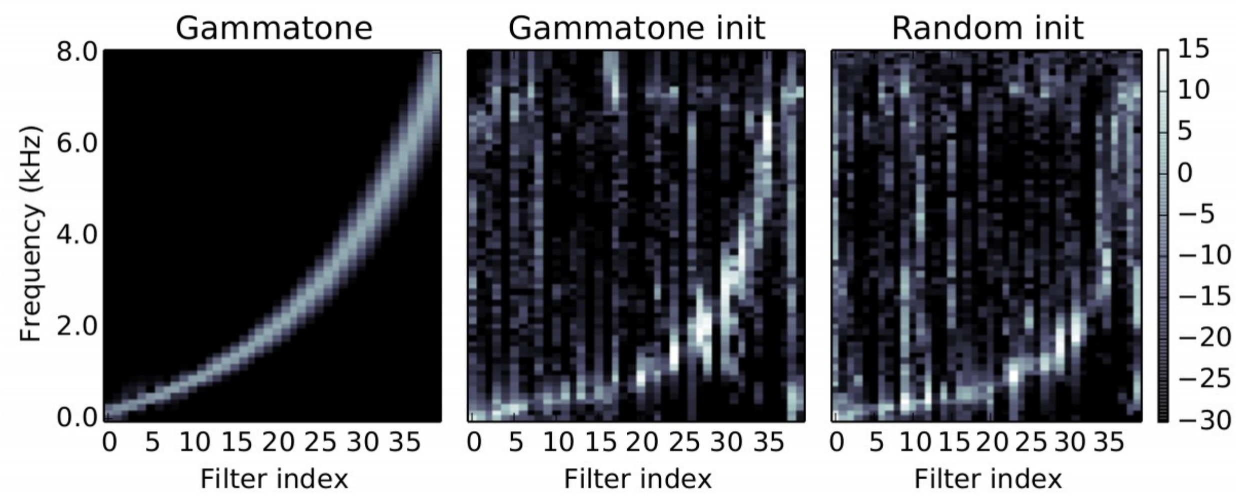
- Log-mel is better (~2.0% abs)
- $\text{WER}_{\text{rand-init}} \approx \text{WER}_{\text{GT-init-train}}$
- $\text{WER}_{\text{GT-init-train}} < \text{WER}_{\text{GT-init-fixed}}$
  - ~ 1.5% abs
- Removing log hurts (~ 1.5% abs)
  - Dynamic range compression is useful

Model	Clean	Fixed	Varied
Mel-fb DNN	25.6%	39.8%	39.5%
Waveform CNN	27.2%	41.6%	41.5%
Waveform CNN mel gammatone fixed	28.8%	43.6%	43.5%
Waveform CNN mel gammatone init	27.1%	41.7%	41.5%
Waveform CNN no log	28.5%	43.0%	42.9%

For single-channel the Fixed and Varied are very similar ...
– same distance, different angle



Learned Filters – Single Channel



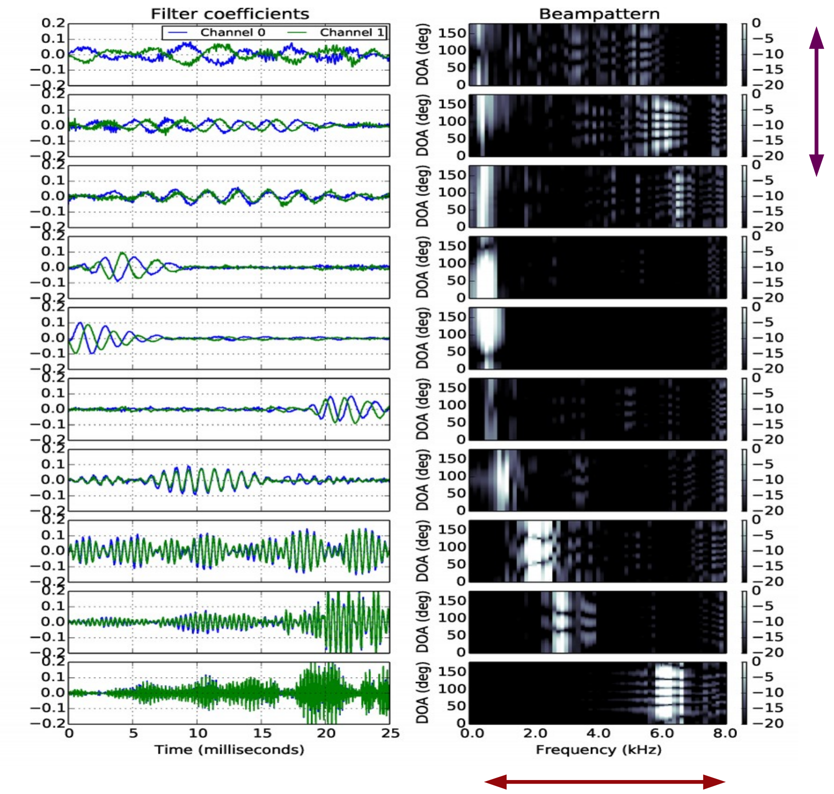
– Centre frequency = argmax {Magnitude Spectrum}

– Loosely auditory-like filters

  * BW increases by f_c & more filters in low frequencies

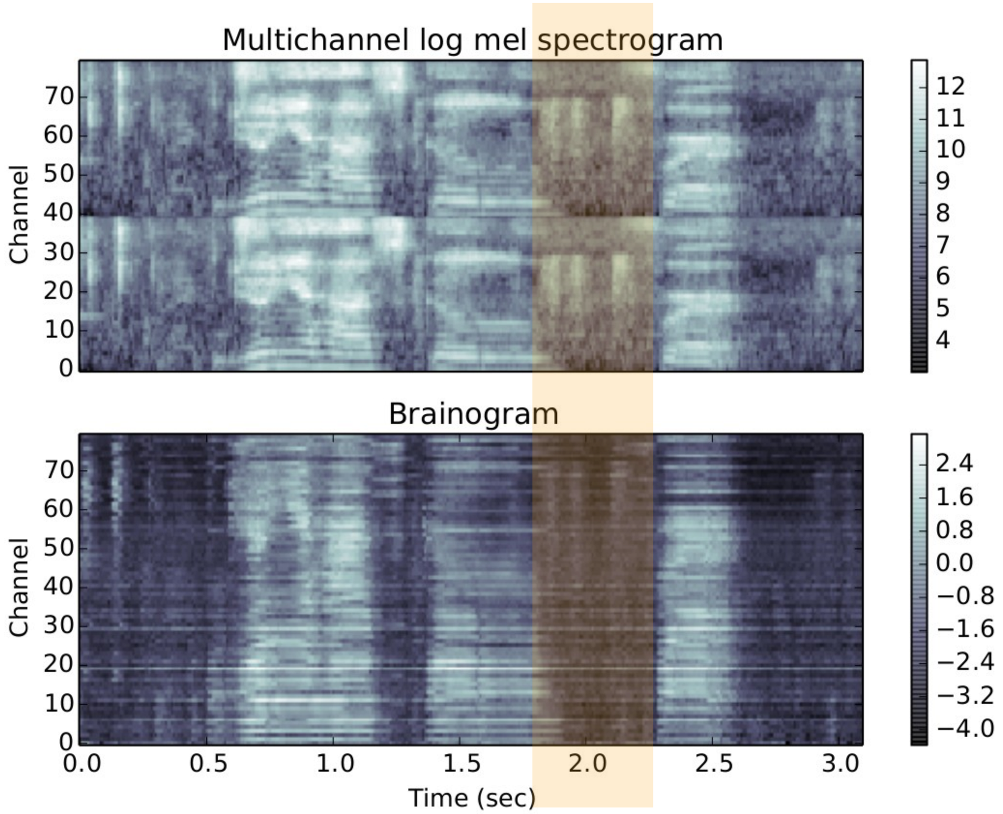
Learned Filters – Multi (2) Channel

- Spatial and spectral filtering
- Learned filters per channel
  - Bandpass
    - Some are multi modal
  - Similar h, different t_d
    - Steer null in noise dir



Learned Filters – Multi Channel

- Brainogram ≡ CRBE
- log-mel vs brainogram
  - Noise suppression ...
    - Steering null



Multi-Channel with Geometry Mismatch

- **Matched** → raw is better than fbank

Model	Train set	Fixed	Varied
Stacked mel-fb DNN	Fixed	39.2%	39.3%
Stacked mel-fb DNN	Varied	39.0%	38.9%
Waveform CNN	Fixed	37.5%	52.0%
Waveform CNN	Varied	38.4%	38.1%
Beamformer mel-fb DNN	Fixed	35.9%	36.6%
Beamformer mel-fb DNN	Varied	36.0%	36.3%

Features: fbank, waveform, fbank+BF
Matched: fixed vs fixed or varied vs varied
Mismatched: fixed vs varied
Beamformer log-mel: Delay-and-Sum

Multi-Channel with Geometry Mismatch

- **Matched** → raw is better
- **Mismatch** → depends ...
  - Fixed-Varied → very poor
  - Var-Fix→better than mel-log

Model	Train set	Fixed	Varied
Stacked mel-fb DNN	Fixed	39.2%	39.3%
Stacked mel-fb DNN	Varied	39.0%	38.9%
Waveform CNN	Fixed	37.5%	52.0%
Waveform CNN	Varied	38.4%	38.1%
Beamformer mel-fb DNN	Fixed	35.9%	36.6%
Beamformer mel-fb DNN	Varied	36.0%	36.3%

– **Features:** **fbank**, **waveform**, **BF+fbank**
– **Matched:** fixed vs fixed or varied vs varied
– **Mismatched:** fixed vs varied
– **BF+fbank**: Delay-and-Sum

Multi-Channel with Geometry Mismatch

- **Matched** → raw is better
- **Mismatch** → depends ...
  - Fixed-Varied → very poor
  - Var-Fix → better than mel-log
- BeamForming helps Mel-fb
  - WER improvement ~ 3-4 %
  - Oracle D+S min mismatch effect

Model	Train set	Fixed	Varied
Stacked mel-fb DNN	Fixed	39.2%	39.3%
Stacked mel-fb DNN	Varied	39.0%	38.9%
Waveform CNN	Fixed	37.5%	52.0%
Waveform CNN	Varied	38.4%	38.1%
Beamformer mel-fb DNN	Fixed	35.9%	36.6%
Beamformer mel-fb DNN	Varied	36.0%	36.3%

– **Features: fbank, waveform, BF+fbank**
– **Matched:** fixed vs fixed or varied vs varied
– **Mismatched:** fixed vs varied
– **BF+fbank**: Delay-and-Sum

SPEAKER LOCATION AND MICROPHONE SPACING INVARIANT ACOUSTIC MODELING FROM RAW MULTICHANNEL WAVEFORMS

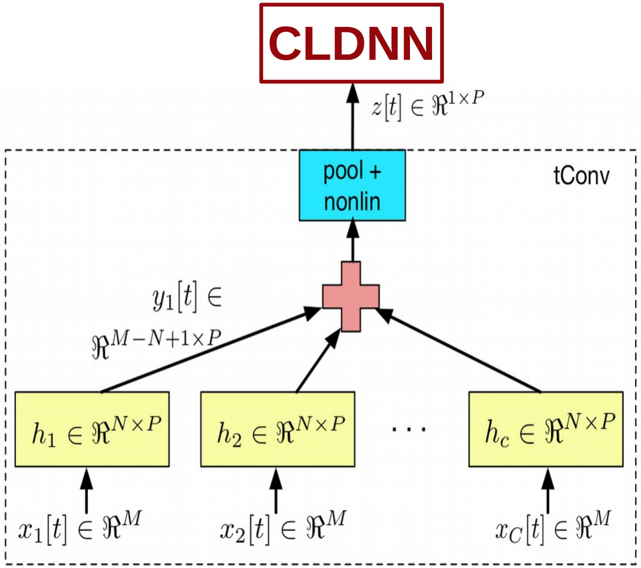
Tara N. Sainath, Ron J. Weiss, Kevin W. Wilson, Arun Narayanan, Michiel Bacchiani and Andrew Senior

Google, Inc., New York, NY, USA
{tsainath, ronw, kwwilson, arunnt, michiel, andrewsenior}@google.com

tConv: Neuro Beamformer (BF)

- No need to localisation
  - Delay absorbed in weights!
- P filters per channel are learned
  - P *look directions* ($\theta_{1:P}$)

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

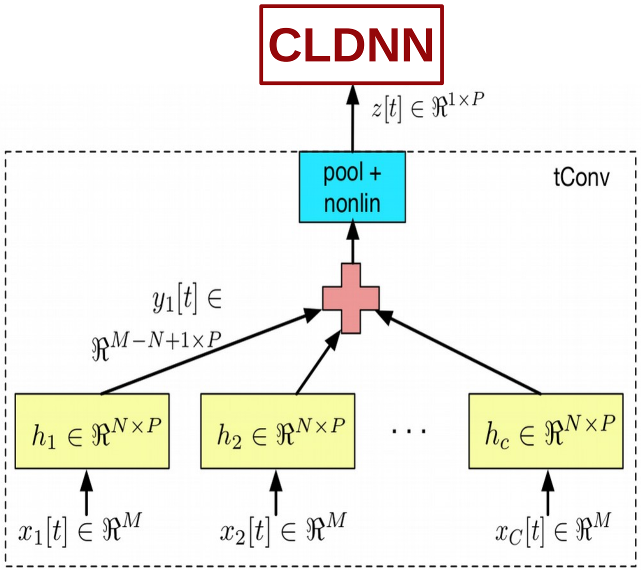


M: input len N: filter len
P: #filters C: #channels

tConv: Neuro Beamformer (BF)

- Parameters
  - #channels_in: C; #channels_out: P
  - #fitler_len: N

- Structure + Size change
  - Conv per ch per p → M-N+1 x P x C
  - **Sum** across channels → M-N+1 x P
  - Max-pooling → 1 x P
  - NonLin → Log(ReLU(.) + 0.01)



M: input length in samples

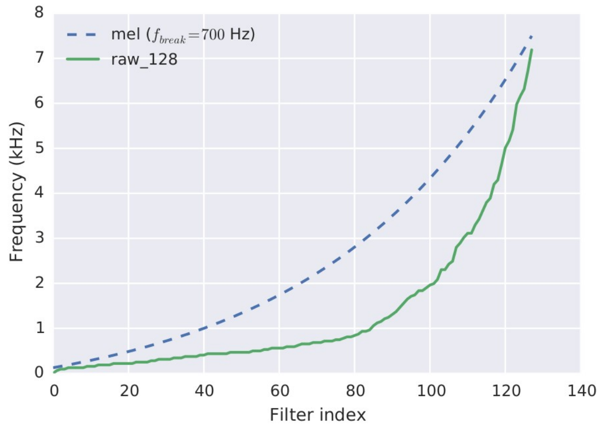
Single-Channel

- Effect of #filters (P)
  - Larger P → lower WER
  - RWERR [40→128]: mel:3.2%; raw: 4.9%

- More filters operating in low frequencies
  - Centre freq ≈ argmax freq response

2000 h, Voice Search

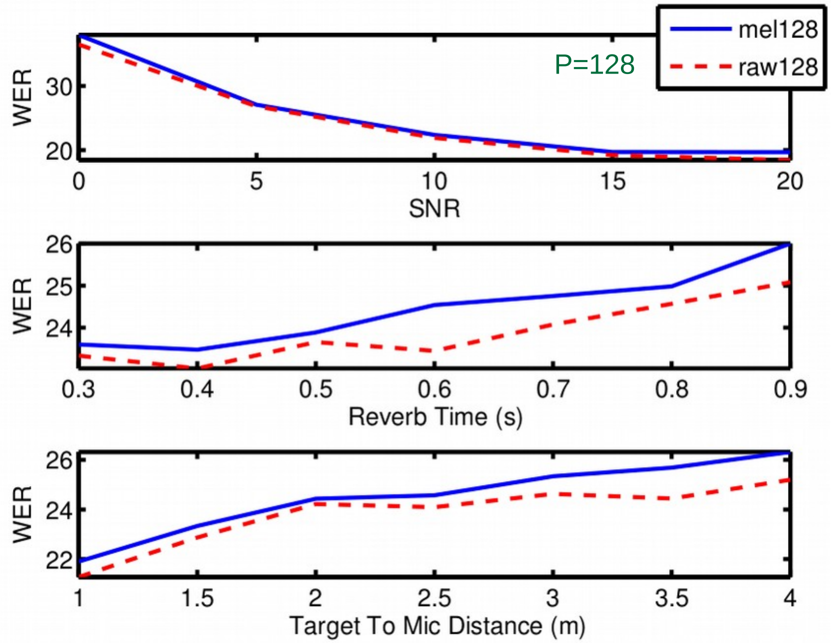
# of Filters (P)	log-mel	raw waveform
40	25.2	24.7
84	25.0	23.7
128	24.4	23.5



M=35ms; N=25ms; P: 128

Single-Channel – Noisy Condition

- Distortion type
  - Additive (SNR) → slightly better
  - Reverberation (T_{60}) → better
  - Far-field (distance) → better

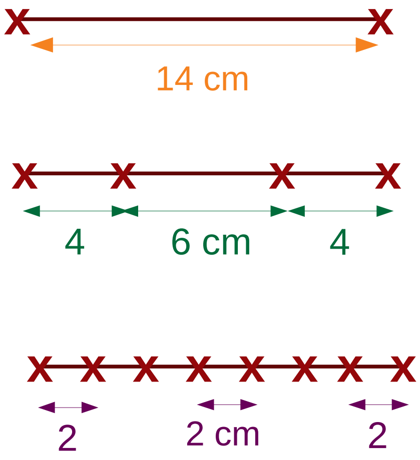


Single vs Multi-Channel

- More channels is better
  - Spatial filtering with higher resolution
- Raw outperforms log-mel
  - Time info
    - Phase spec → delay est

Multi-channel

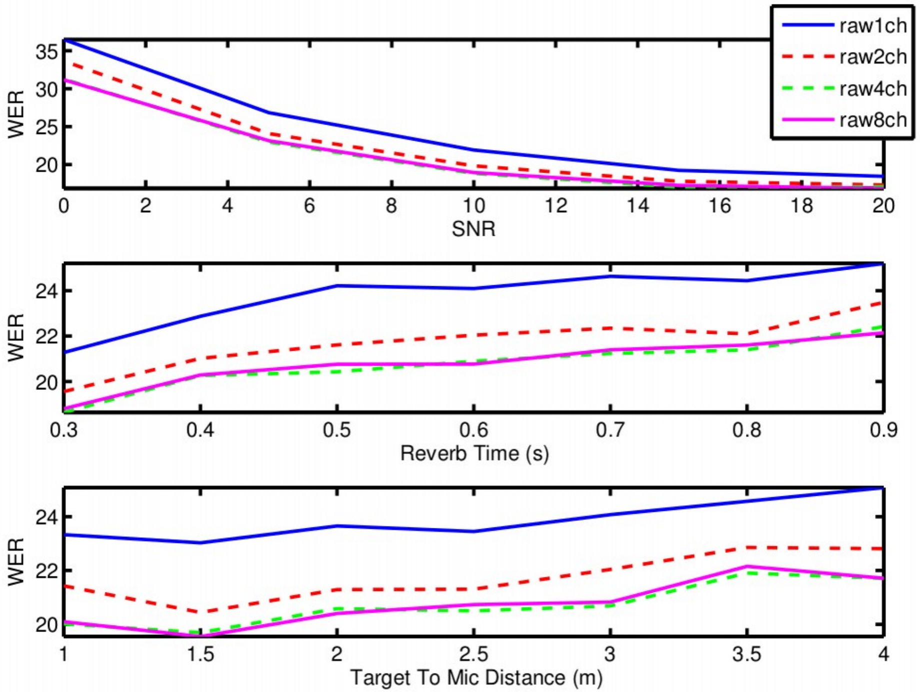
Feature	1ch	2ch (14cm)	4ch (4-6-4cm)	8ch (2cm)
log-mel	24.4	22.0	21.7	22.0
raw	23.5	21.8	21.3	21.1



Single vs Multi-Channel

- More channels is better
  - Saturation after 4

- Distortion type
  - Additive: 2-6% better (abs)
  - Reverberation (T_{60}) → 2-3%
  - Far-field (distance) → ~ 3%



Comparison with Oracle Experiments

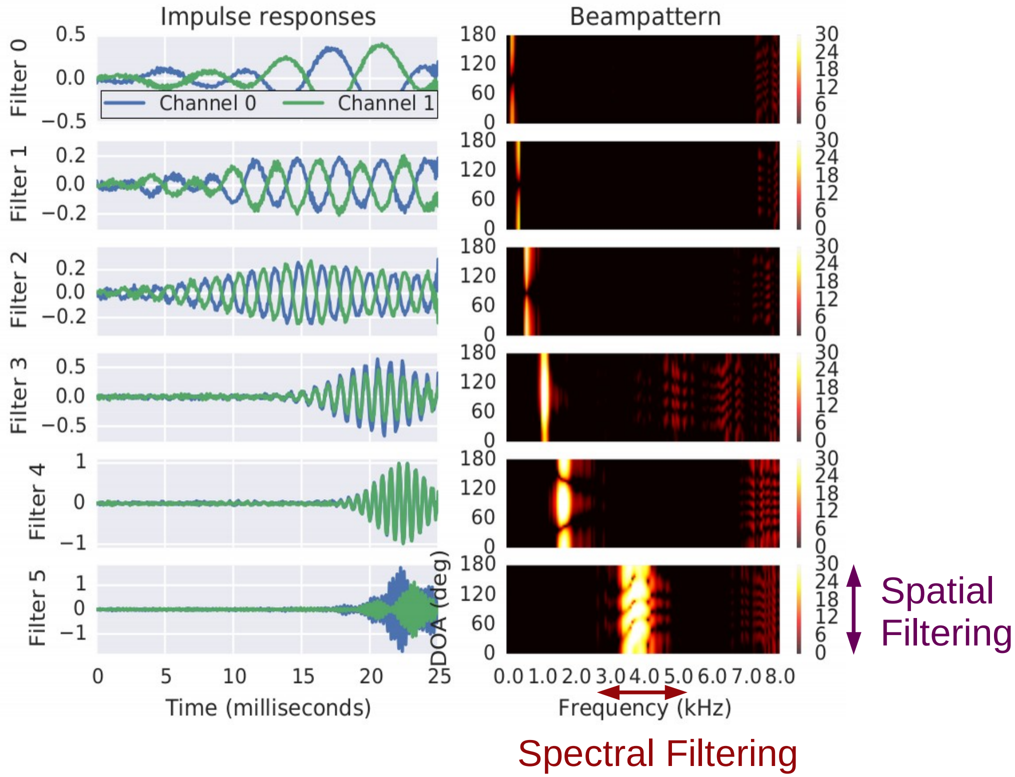
- Raw waveform systems:
  - D+S: Align w/ oracle delay → sum → single-channel
  - TAM: Align w/ oracle delay → multi-channel
  - Raw, no tdoa: neuro-beamforming

- Neuro-BF w/o localisation
  - Outperforms oracle D+S!
  - Similar to TAM
    - Delay is not important!

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, no tdoa	23.5	21.8	21.3	21.1

Learned Filters – 2-channel

- Simultaneous Spatial and Spectral filtering
- Ch_0 vs Ch_1 Similar & delayed
  - Steering a null
    - Delay ≡ Direction
- Larger BW for high f_c



Geometric Mismatch; 2-channel

- Geometric mismatch experiments
  - Trained on 14 cm mic spacing
  - Test with 14, 10, 6, 2 cm

Method	14cm	10cm	6cm	2cm
raw, 1ch	23.5	23.5	23.5	23.5
D+S, 2ch, tdoa	22.8	23.2	23.3	23.7
TAM, 2ch, tdoa	21.7	22.1	23.2	30.6
raw, 2ch	21.8	22.2	23.3	30.7

- D+S → stable performance
  - Does not see the mismatch
- TAM & raw → stable except for 2cm
  - Handle reasonable mismatch
  - Train on 14, train on 2cm → strong mismatch

Multi-Geometric Training (MGT)

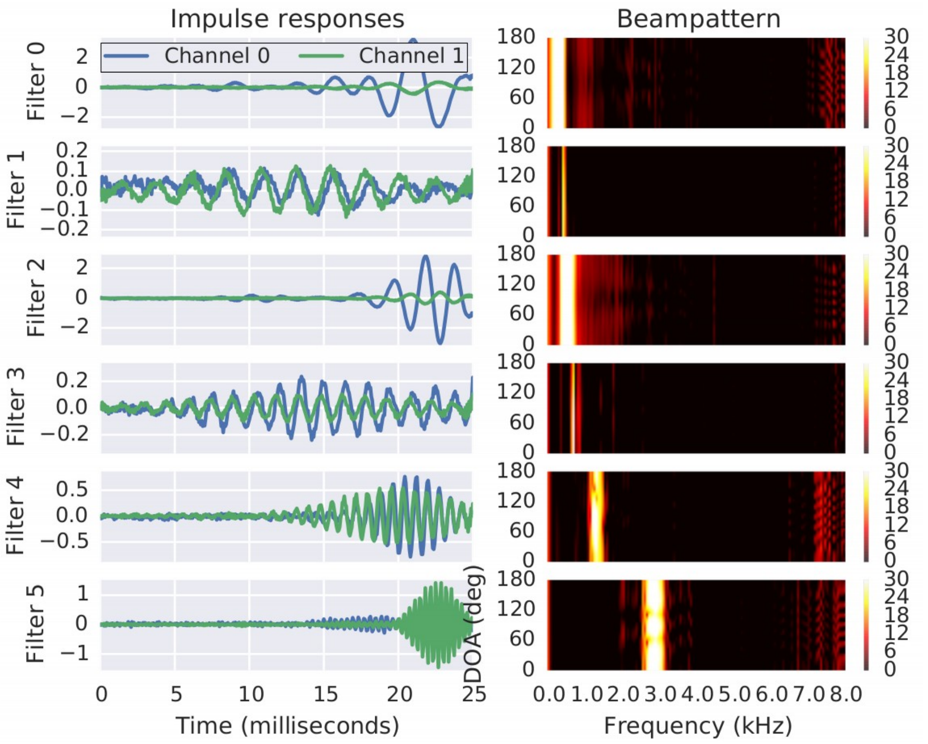
- MGT to deal with geometric mismatch
- System well handles 2-14cm spacing
- Works w/o delay knowledge!
  - Outperforms single-channel!

Method	14cm	10cm	6cm	2cm
raw, 2ch	21.8	22.2	22.3	30.7
raw, 2ch, multi-geo	21.9	21.7	21.9	21.8

Method	1ch, repeated twice
raw, 1ch	23.5
raw, 2ch	33.9
raw, 2ch, multi-geo	23.1

Learned Filters – Multi-Geo

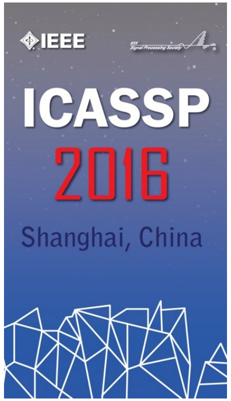
- No longer exhibit strong spatial response
  - $\max@f_c - \min@f_c > 6\text{dB}$
  - No null
- Larger BW for high f_c ?



FACTORED SPATIAL AND SPECTRAL MULTICHANNEL RAW WAVEFORM CLDNNS

Tara N. Sainath, Ron J. Weiss, Kevin W. Wilson, Arun Narayanan, Michiel Bacchiani

Google, Inc., New York, NY, USA
{tsainath, ronw, kwwilson, arunnt, michiel}@google.com



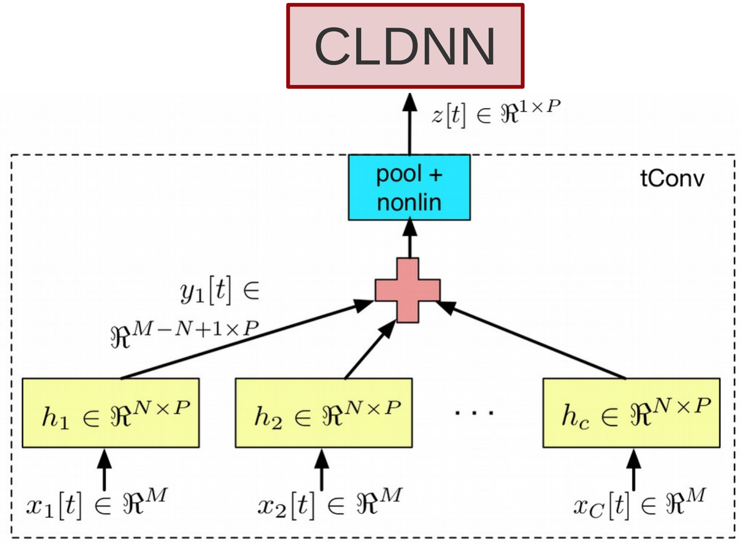
Factorised Neuro-Beamformer

- **Unfactorised**
  - tConv: Simultaneous spatial and spectral filtering
  - Pooling + Non-linearity
- **Factorised**
  - Intuition: Factor out spectral and spatial filtering
    - tConv1 → spatial
    - tConv2 → spectral
    - Pooling + Non-linearity

Factorised Neuro-Beamformer

- **Unfactorised**
  - tConv: Simultaneous spatial and spectral filtering
  - Pooling + Non-linearity
- **Factorised**
  - Intuition: Factor out spectral and spatial filtering
    - tConv1 → spatial & spectral
    - tConv2 → spectral
    - Pooling + Non-linearity

Factorised vs Unfactorised Model

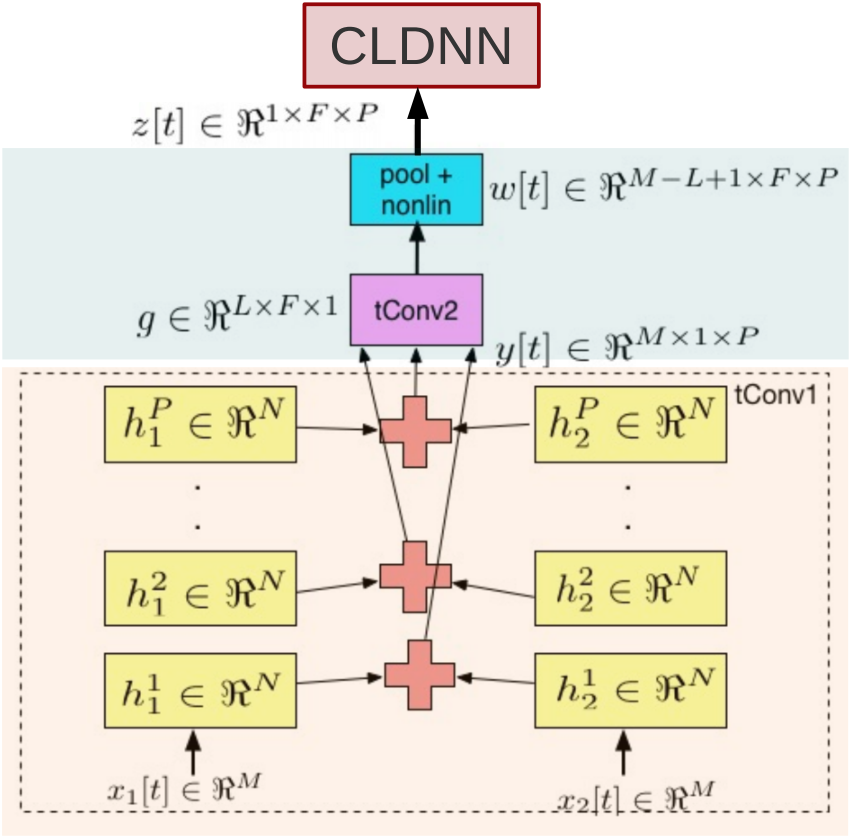


Unfactorised: simultaneous
Spectral and spatial filtering

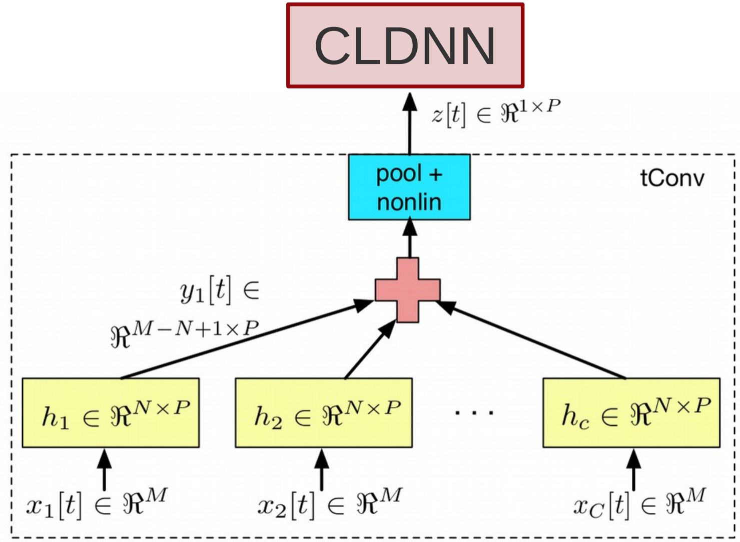
Factorised vs Unfactorised Model

Spectral Filtering (tConv2)

Spatial Filtering (tConv1)

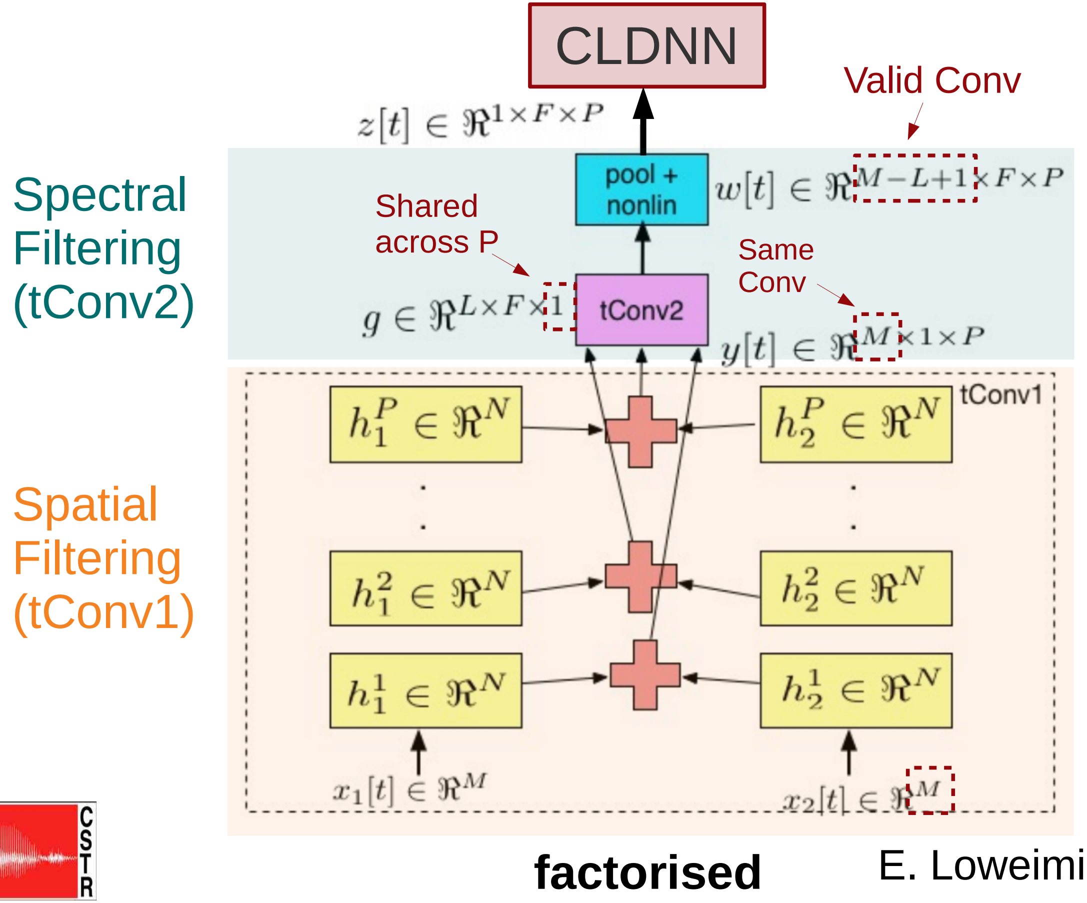


factorised



Unfactorised: simultaneous Spectral and spatial filtering

Factorised Model – Parameters



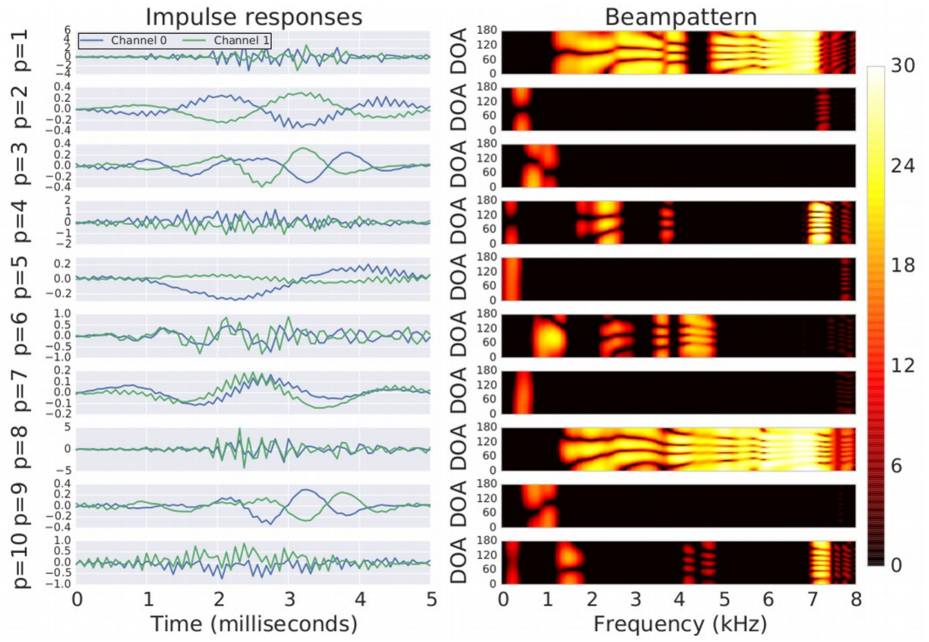
- M: input length (560)
- C: #channels (mic) = 2
- N: tConv1 filter len (80)
- P: tConv1 #filters
- g: tConv2
- L: tConv2 filter len (400)
- F: tConv2 #filters (128)
- NonLin: log (ReLU(.) + 0.01)

Factorised vs Unfactorised Model

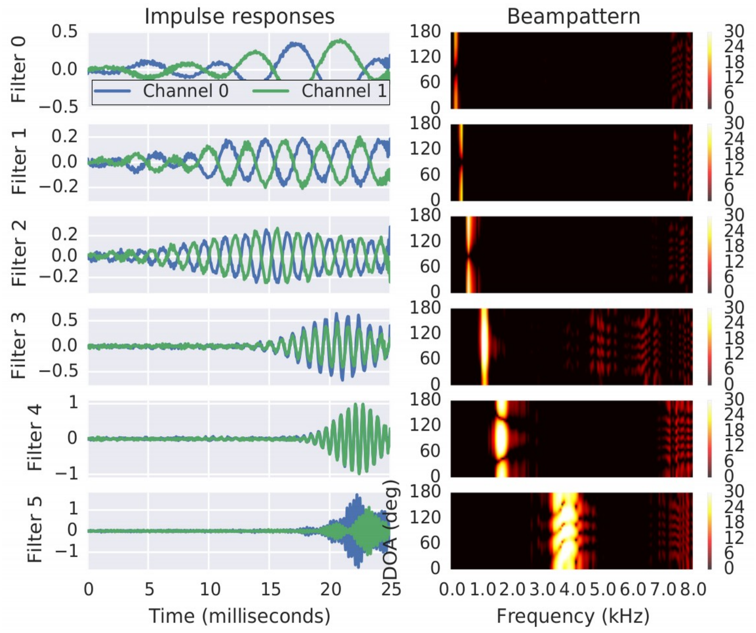
- tConv1: M = 560 samples [35ms], N=80 [5ms], tConv2: F=128, L=400,
- Filter size (N) and #filters (P) is much smaller in *tconv1*
  - Small N → broadband response → less spectral resolution
  - Small P → a few spatial look directions
- *Nonlin+pooling* ONLY after tconv2, not tconv1
- *Same* vs *Valid* convolution types
- Feature dim: $x_t \in R^{1xFxP}$ vs $x_t \in R^{1xP}$
- *tConv2*
  - Longer-duration (better freq resolution), Single-channel filters
  - Filters $\in R^{LxFx1}$, shared across P input feature maps
  - Convolution type: Valid => output $R^{M-L+1xFxP}$ → pooling → $x_t \in R^{1xFxP}$

Factorised vs Unfactorised Model

Factorised



Unfactorised



– Spatial Behaviour
  * wider beams + strong spatial response
  * steering null

– Spectral Behaviour
  * multi-modal with different BWs

Experimental Results

- Factored vs Unfactored
  - 6.4% RWERR
- Higher P $\rightarrow$ lower WER
  - P $\leq$ 10 Comp. Complexity
- tConv1
  - Trained vs fixed:
    - 4.6% RWERR

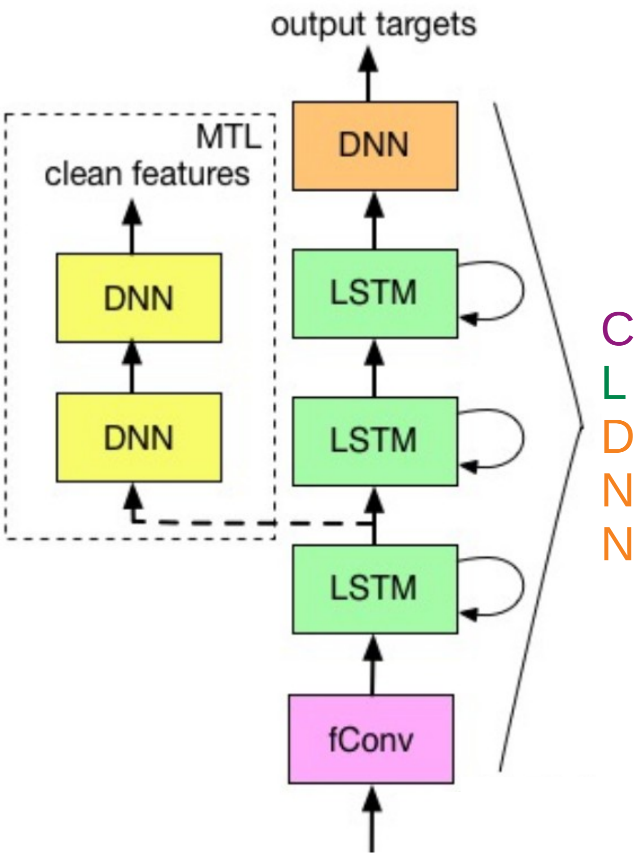
# Spatial Filters P	WER
baseline 2 ch, raw [1]	21.8
1	23.6
3	21.6
5	20.9
10	**20.4**

# Spatial Filters P	`tConv1` Layer	WER
5	fixed	21.9
5	trained	20.9

Fixed: oracle D+S

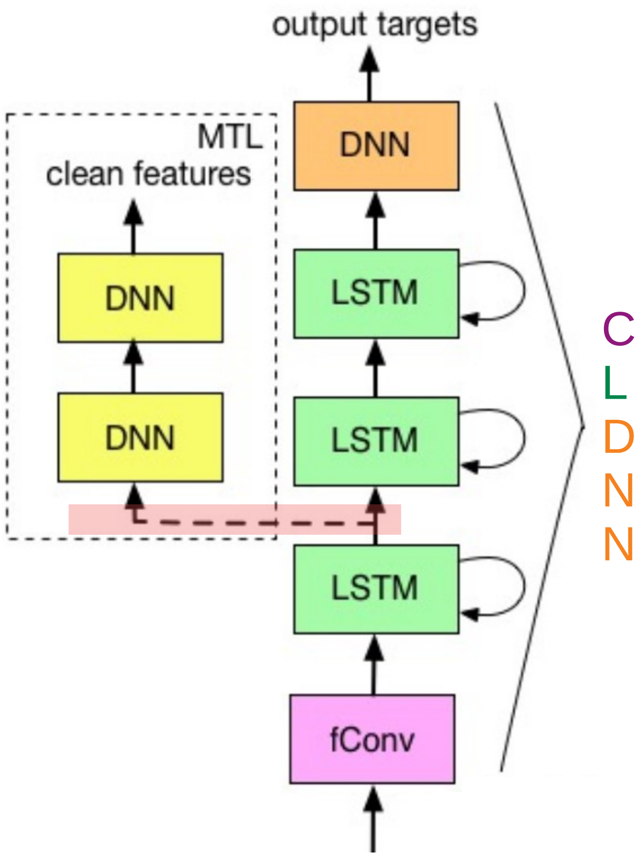
MTL on Unfactored Model

- Speech Enhancement via DNNs
  - Auto-Enc, TF mask, MTL, etc.
- **Multi**-**T**ask **L**earning (MTL)
  - **ASR** predicts CD states
  - **Denoising** predicts clean log-mel
  - Loss = α CE_{ASR} + (1-α) MSE_{Enh}
    - Here, α = 0.9



MTL on Unfactored Model

- Speech Enhancement via DNNs
  - Auto-Enc, TF mask, MTL, etc.
- **Multi**-**T**ask **L**earning (MTL)
  - **ASR** predicts CD states
  - **Denoising** predicts clean log-mel
  - Loss = α CE_{ASR} + (1-α) MSE_{Enh}
  - **Optimal branch position ???**



MTL Optimal Position

- Higher layers better!
  - After 1LSTM or DNN is optimal
  - Why?
- Max gain (RWERR)
  - 1 Ch → 3.8%
  - 2 Ch → 5.0 %

Unfactored model

Denoising task branching layer	1 channel	2 channel
no MTL [1]	23.5	21.8
tConv	23.2	21.7
fConv	23.2	21.8
1LSTM	22.6	20.7
DNN	22.6	20.7

* MTL after …

-- tConv
-- fConv
-- 1LSTM: 1st LSTM layer
-- DNN (just before output layer)

Training → CE vs Seq

- D+S → oracle delay
- MVDR → oracle delay and noise/speech cov mat
  - Optimal in SINR
- Neuro-BF outperforms MVDR!
- MTL gain for factored: 2%

2 Ch, P=10

Method	CE	Seq
log-mel, 1 channel	25.2	20.7
raw, 1 channel	23.5	19.2
delay-and-sum, 8 channel	22.4	18.8
MVDR, 8 channel	22.4	18.7
unfactored raw, 2 channel [1]	21.8	18.2
factored raw, 2 channel	20.4	17.3
factored raw, 2 channel, MTL	**20.0**	**17.0**

Conclusion – Part 3

- Raw waveform outperforms log-mel in CLDNN AM
  - On 2000 h; min data amount for better perfomance?
- Neuro-beamforming
  - w/o localisation, outperforms MVDR with oracle info
  - Unfactorised: simultaneous Spectral & Spatial filtering
  - Factorised: dissociates spectral and spatial filtering
- MTL → ASR + Enhancement → branch at high levels → helps

That’s It!

- Thanks for Your Attention!
- Q & A

- Next Session:
  - Parametric CNNs for Raw waveform AM