On the Robustness and Training Dynamics of Raw Waveform Models

Erfan Loweimi
Peter Bell and Steve Renals

CSTR Talk
10,May, 2021

On the Robustness and Training Dynamics of Raw Waveform Models

Erfan Loweimi, Peter Bell and Steve Renals

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

{e.loweimi, peter.bell, s.renals}@ed.ac.uk

Rejected in ICASSP 2020

Accepted in INTERSPEECH 2020

On the Robustness and Training Dynamics of Raw Waveform Models

Erfan Loweimi, Peter Bell and Steve Renals

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

{e.loweimi, peter.bell, s.renals}@ed.ac.uk

Rejected in ICASSP 2020

Accepted in INTERSPEECH 2020

Life is not fair … Never give up!

Outline

- Raw waveform acoustic modelling
- Dynamics
- Robustness
- Conclusion

Outline

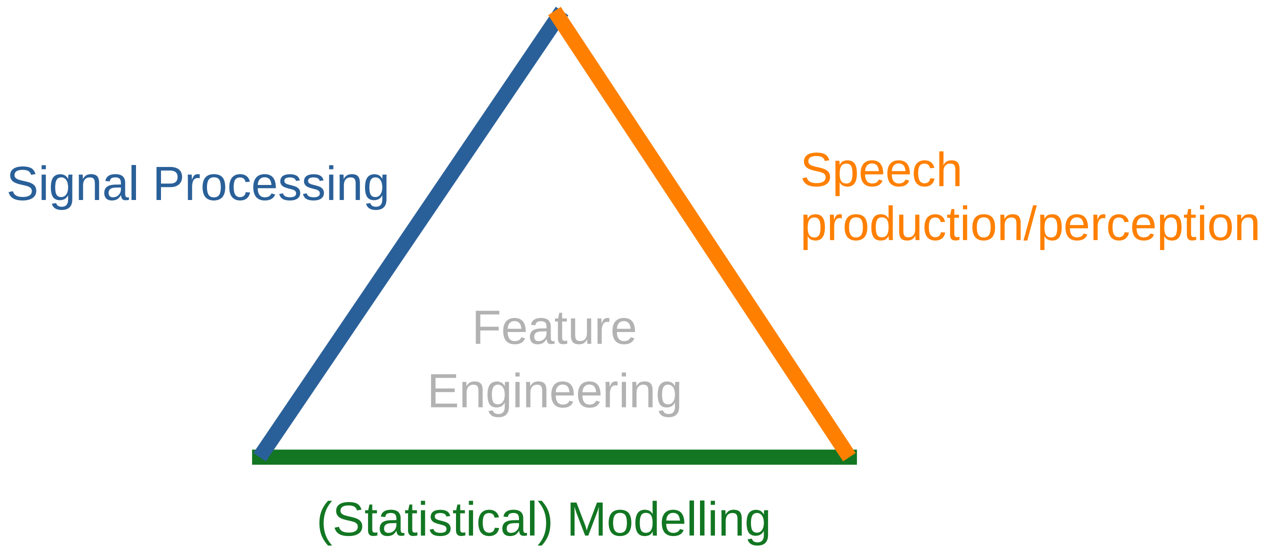
- Raw waveform acoustic modelling
  - Feature engineering vs learning
  - Pros & cons
- Dynamics
- Robustness
- Conclusion

Feature Engineering: Goal

- Goal: A handcrafted pipeline

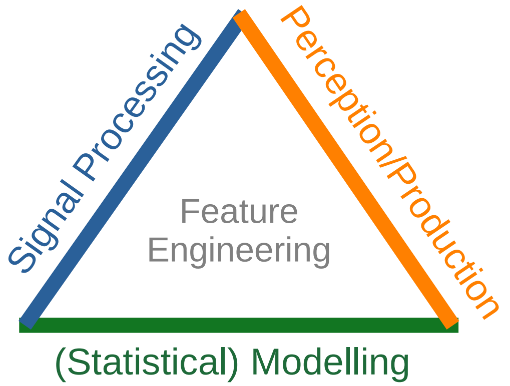
Feature Engineering: Design

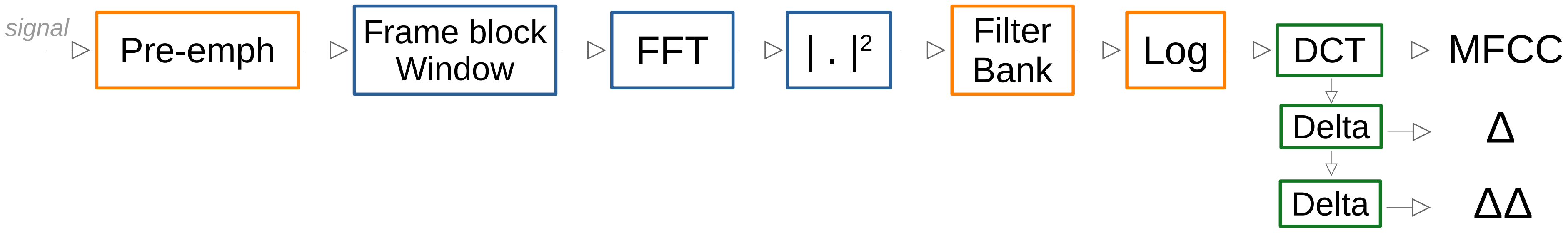
- Design: Prior knowledge ...



Feature Engineering: Design

- Design: Prior knowledge ...





Feature Engineering: Pros

- Pros: Interpretable, easy, fast, general-purpose

Feature Engineering: Pros

- Pros: Interpretable, easy, fast, general-purpose

MFCC is successfully used in many tasks ...

ASR

TTS

Speaker ID

Emotion classification

Language ID

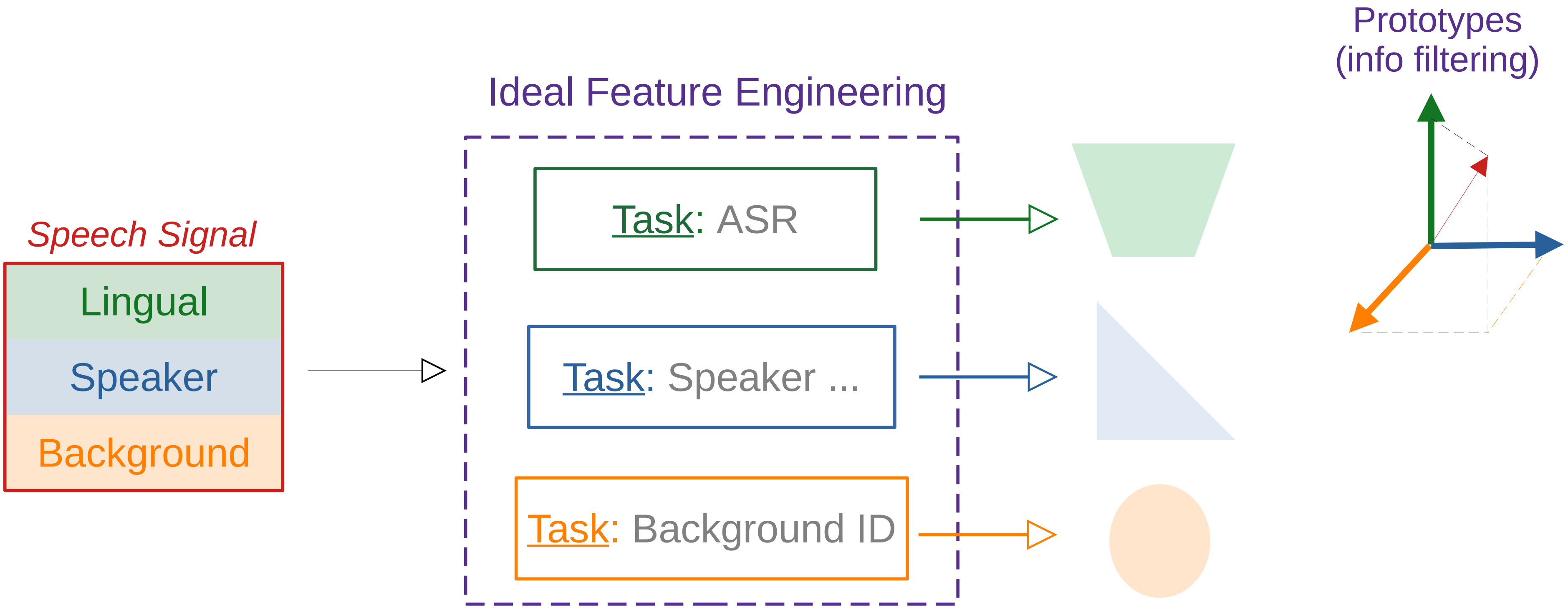
and many more ...

Feature Engineering: Cons (1)

- Task-blind (general-purpose)

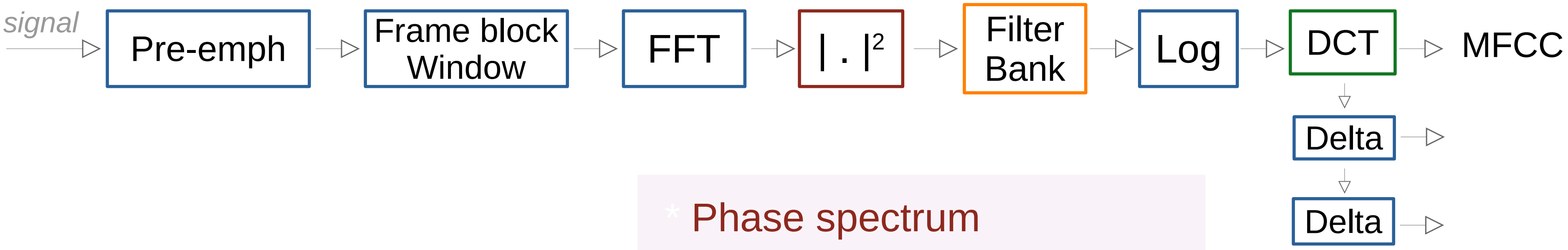
Feature Engineering: Cons (1)

- Task-blind (general-purpose)



Feature Engineering: Cons (2)

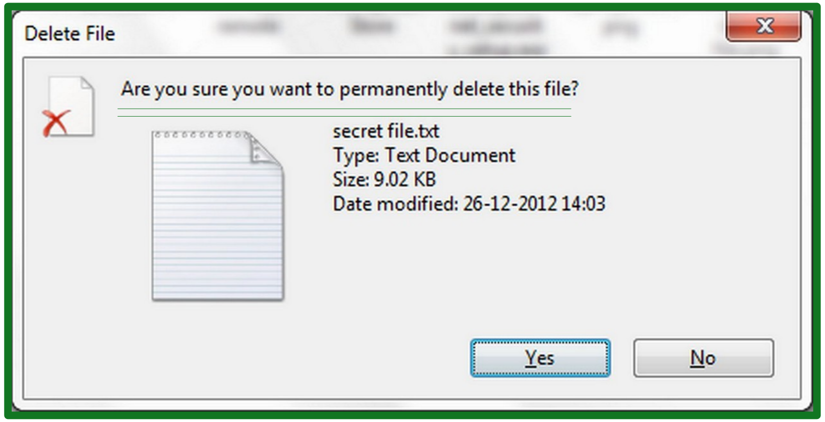
- Suboptimal info loss



- Phase spectrum
- Resolution (subsampling)
- Speaker … (Low-pass lifter)

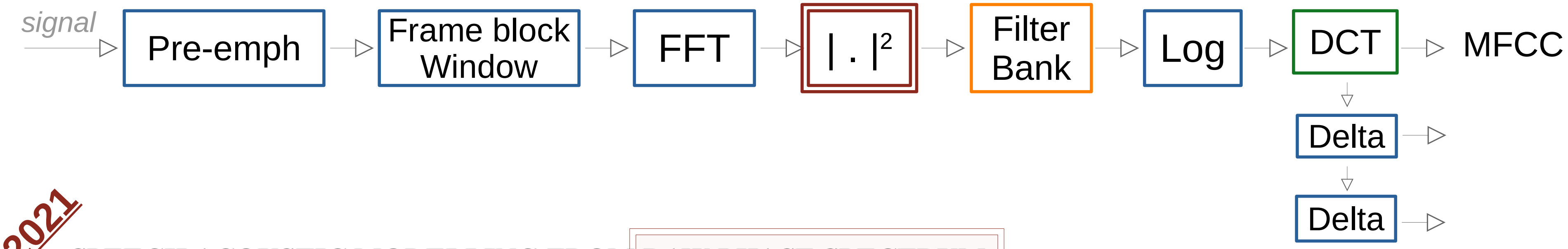
Feature Engineering: Cons (2)

- Suboptimal info loss
  - Lost info is lost permanently



Feature Engineering: Cons (2)

- Suboptimal info loss



ICASSP 2021

SPEECH ACOUSTIC MODELLING FROM RAW PHASE SPECTRUM

Erfan Loweimi [1], Zoran Cvetkovic [2], Peter Bell [1] and Steve Renals [1]

[1] Centre for Speech Technology Research (CSTR), University of Edinburgh, UK
[2] Department of Engineering, King's College London, UK

Feature Engineering: Cons (3)

- Suboptimal info filtering

Link

Optimal Info Filtering: Pass through ONLY relevant/useful info

Feature Engineering: Cons (3)

- Suboptimal info filtering
  - Irrelevant/nuisance info/variability passed through

Link

Feature Engineering: Cons (2) & (3)

- Suboptimal info loss/filtering
  - Lost info is lost permanently
  - Irrelevant/nuisance info/variability passed through

*… The useful information which is not passed to the ASR system is **lost forever**. On the other hand, **irrelevant information** which is not removed has to be dealt with by the ASR system, often at **significant expense**.*

Hermansky et al., “Perceptual Properties of Current Speech Recognition Technology”, Proceedigs of eht IEEE, 2013

Feature Engineering: Cons (2) & (3)

- Suboptimal info loss/filtering
  - Lost info is lost permanently
  - Irrelevant/nuisance info/variability passed through

Speech Acoustic Modelling using Raw Source and Filter Components

Erfan Loweimi [1], *Zoran Cvetkovic* [2], *Peter Bell* [1], *and Steve Renals* [1]

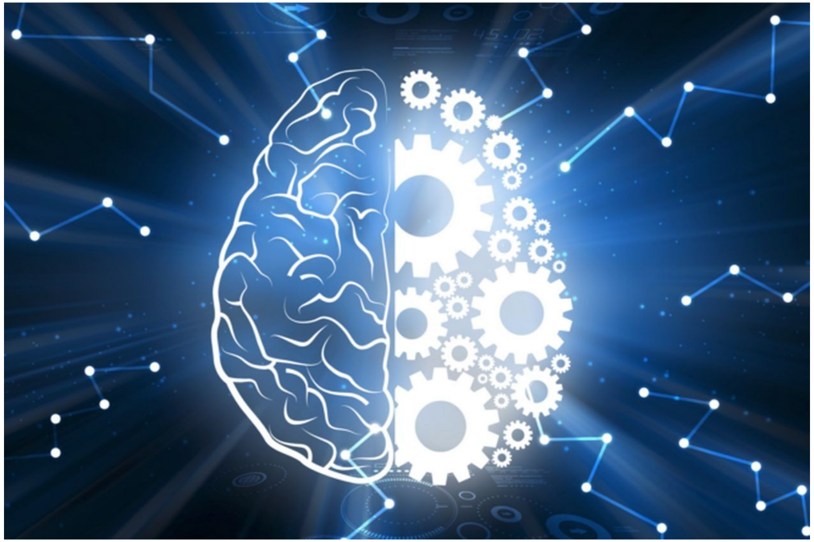
[1] Centre for Speech Technology Research (CSTR), University of Edinburgh, UK
[2] Department of Engineering, King's College London, UK
{e.loweimi, peter.bell, s.renals}@ed.ac.uk zoran.cvetkovic@kcl.ac.uk

Submitted to INTERSPEECH 2021

… task-irrelevant info could be useful **if** ...

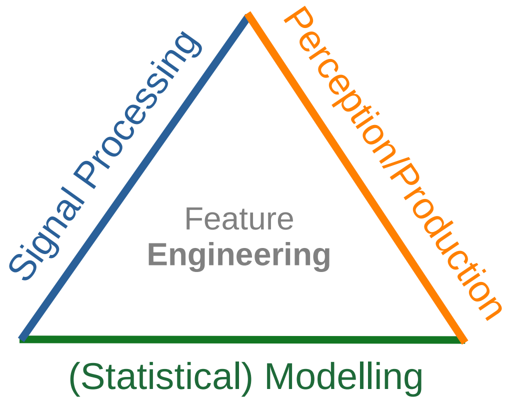
Feature Learning: Goal

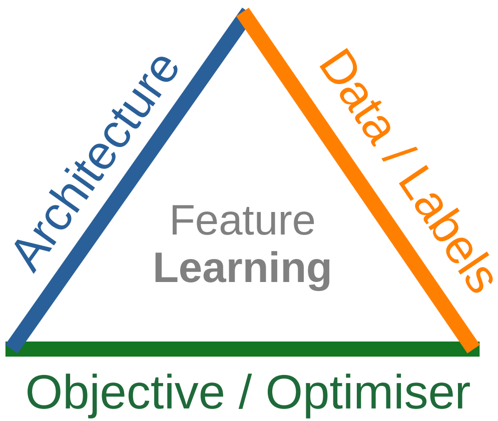
- Goal: Learn the pipeline, instead of engineering



Feature Learning: Design

- Design: Architecture, Data/Labels, Objective/Optimiser





Feature Learning: Pros (1)

- Pros: Task-specific, ~~general-purpose~~ ...

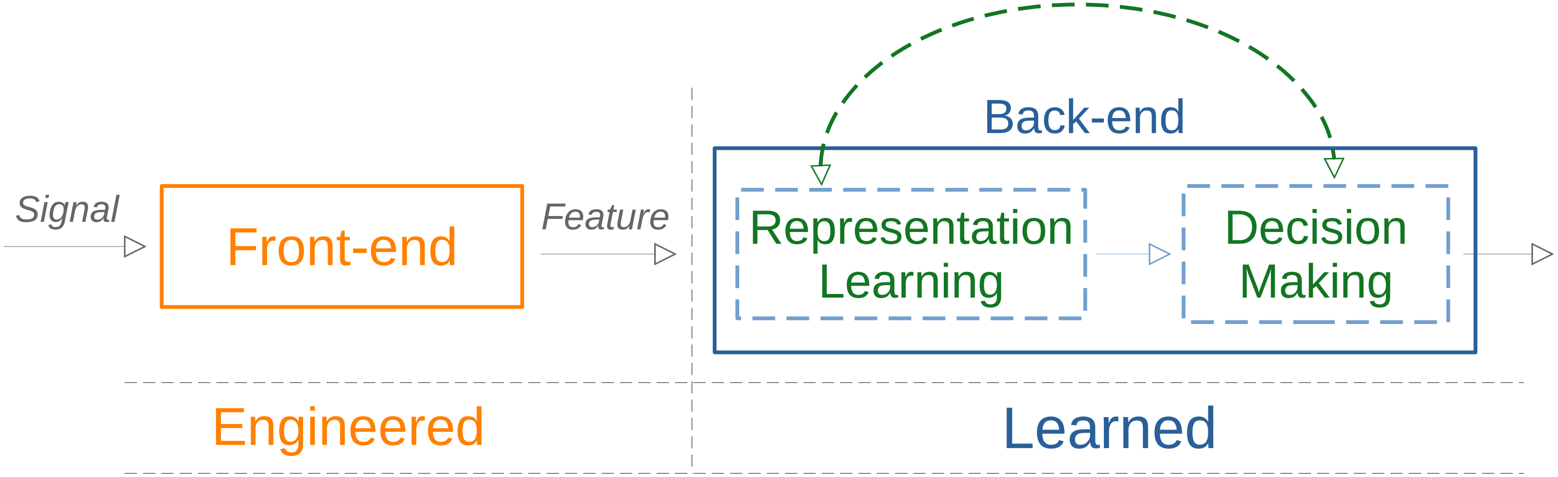
Feature Learning: Pros (1)

- Pros: Task-specific, ~~general-purpose~~ ...

. . .

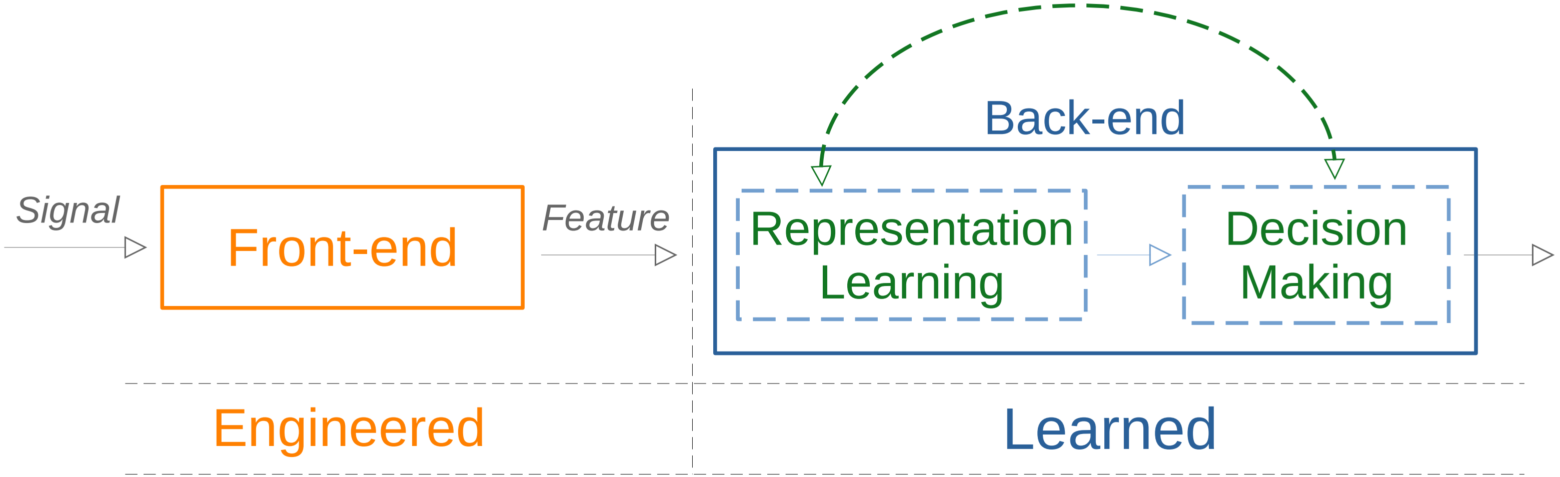
Feature Learning: Pros (2)

- Pros: Joint learning



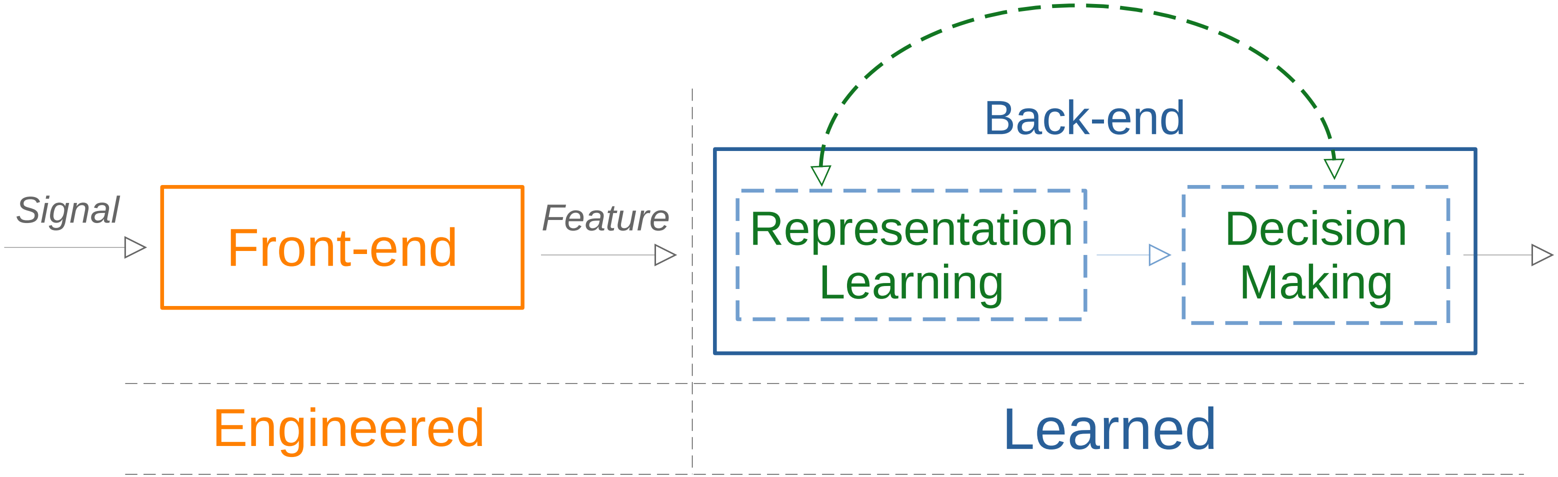
Feature Learning: Caveat

- Info lost in engineering stage is lost permanently ...



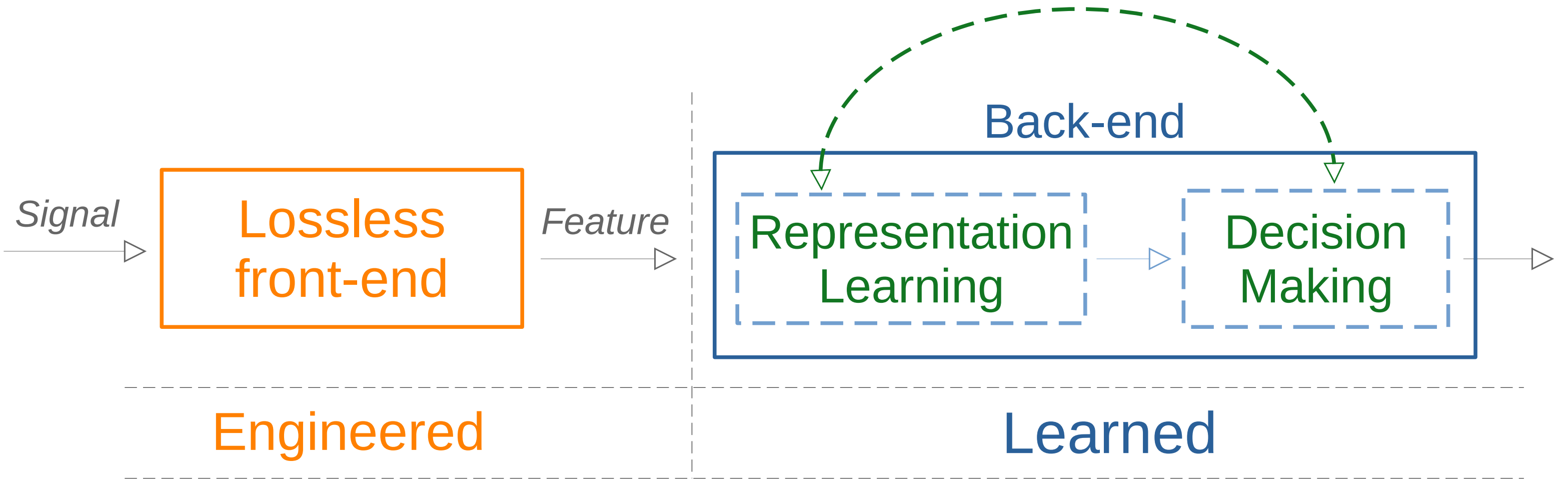
Feature Learning: Caveat

- Info lost in engineering stage is lost permanently ...
  - upperbounds performance
  - machinery cannot generate info



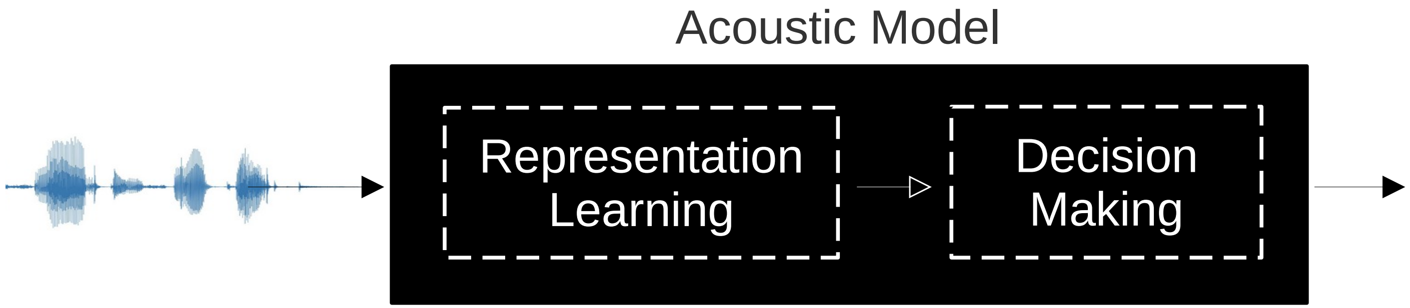
Feature Learning – Caveat Solution

- **Lossless** front-end (signal is uniquely recoverable from feature)
  - Examples: Raw waveform, Mag+Sign, ...



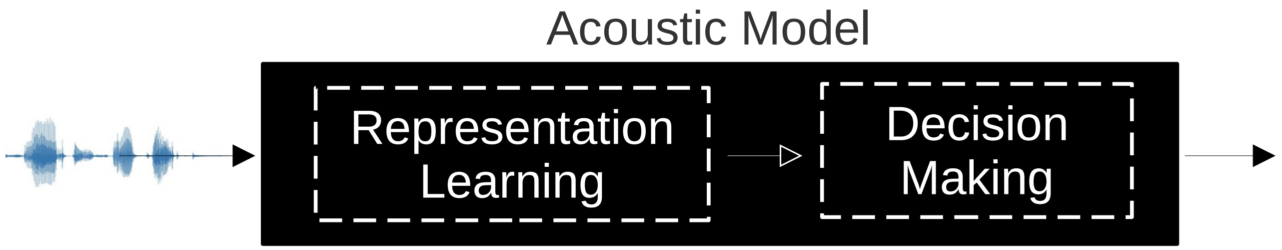
Raw Waveform Acoustic Modelling

- Feed the model with raw waveform



Raw Waveform Acoustic Modelling

- **Pros**:
  - Lossless front-end
  - Task-specific
  - Joint optimisation
  - Interpretability



Raw Waveform Acoustic Modelling

- **Cons**:
  - High-dim … hardware + curse of dimensionality (?)
  - Info disentanglement is challenging
  - Task-specific
  - …

Raw Waveform Acoustic Modelling

- **Solutions**:
  - Data ↔ High-dim + info disentanglement
  - Constraint (arch., regular./norm) ↔ High-dim
  - Adaptation ↔ Task-specific
  - ...

Raw Waveform Acoustic Modelling

- **Solutions**:
  - Data ↔ High-dim + info disentanglement
  - Constraint (arch., regular./norm) ↔ High-dim
  - Adaptation ↔ Task-specific

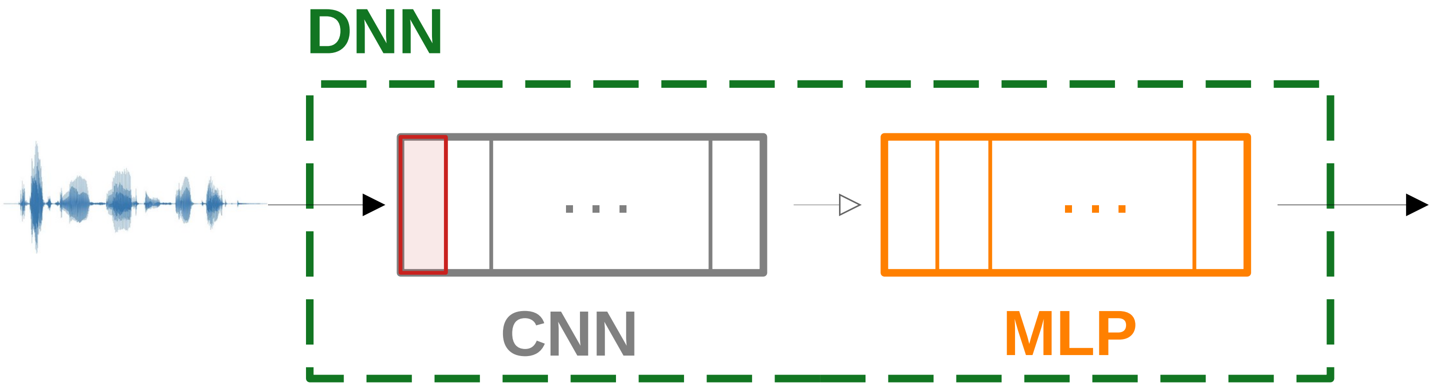
ACOUSTIC MODEL ADAPTATION FROM RAW WAVEFORMS WITH SINCNET

Joachim Fainberg, Ondřej Klejch, Erfan Loweimi, Peter Bell, Steve Renals

Centre for Speech Technology Research, University of Edinburgh, United Kingdom

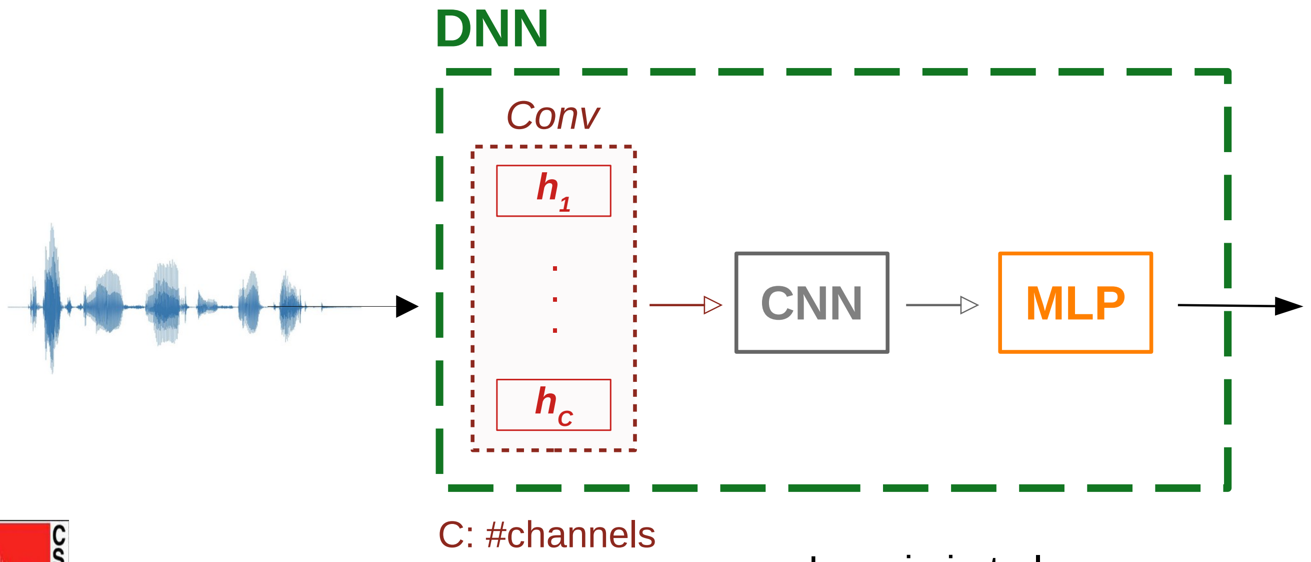
Raw Waveform Acoustic Modelling

- **Pros**: … Interpretability ...



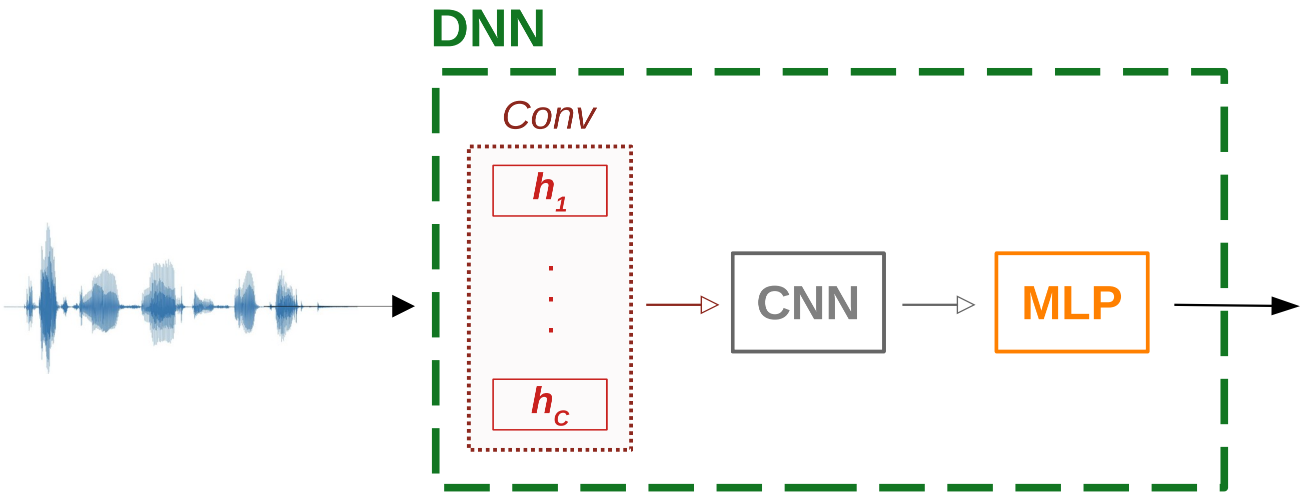
Raw Waveform Acoustic Modelling

- **Pros**: ... Interpretability ...
  - First layer in CNN → Filterbank → Time-Frequency Analysis (TFA)



Raw Waveform Acoustic Modelling

- **Pros**: … Interpretability …
  - First layer in CNN → Filterbank → TFA

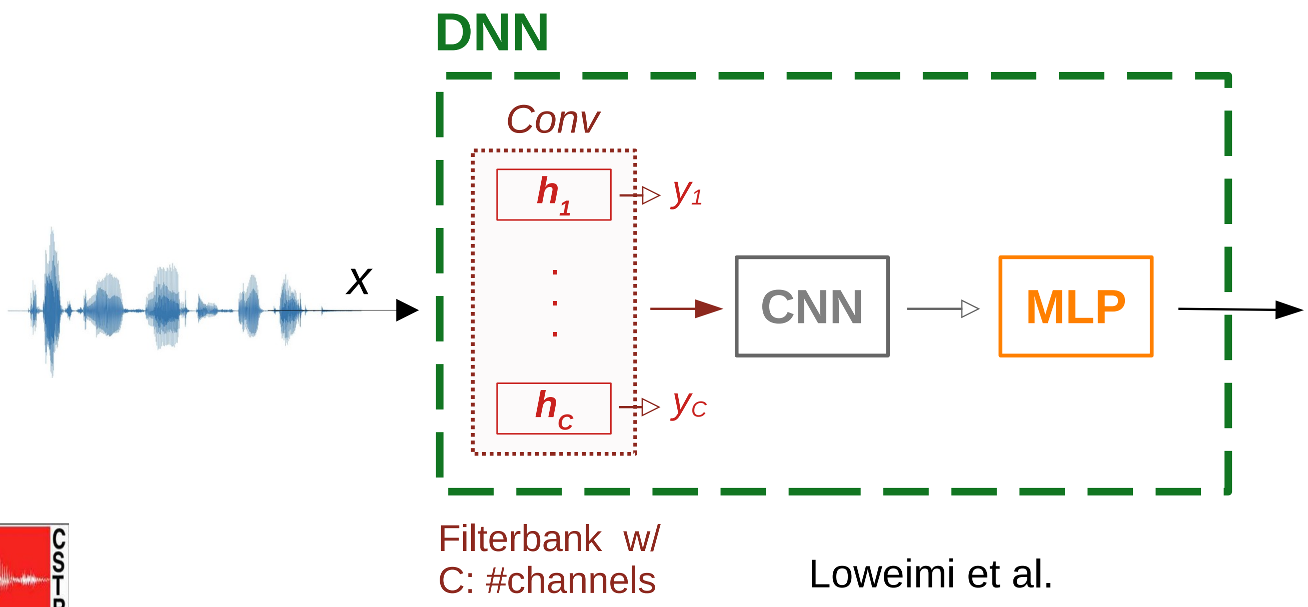


$h_i[t]$: impulse response of i^{th} filter
$H_i[t]$: frequency response of i^{th} filter
Filterbank: $\{h_i \mid 1 \le i \le C\}$

Filterbank w/
C: #channels

Raw Waveform Acoustic Modelling

- **Pros**: … Interpretability ...
  - First layer in CNN → Filterbank → TFA

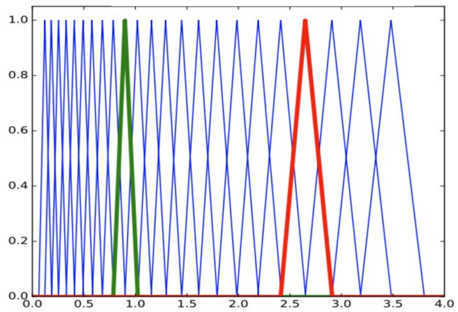


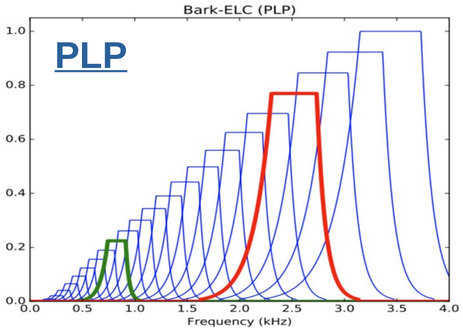
$$y_i(t) = x(t) * h_i(t)$$
$$Y_i(\omega) = X(\omega)\ H_i(\omega)$$

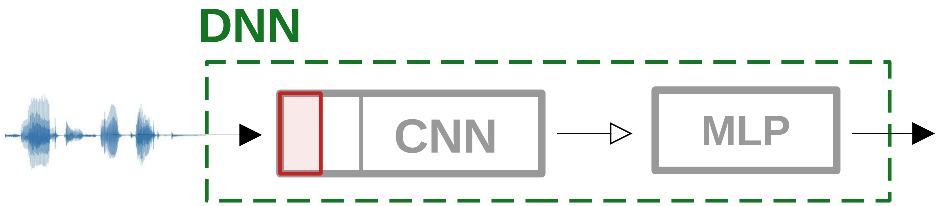
Loweimi et al.

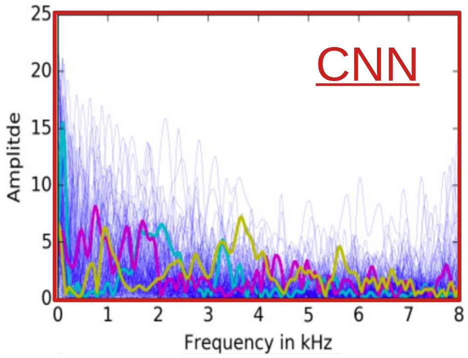
Engineered vs Learned Filterbank

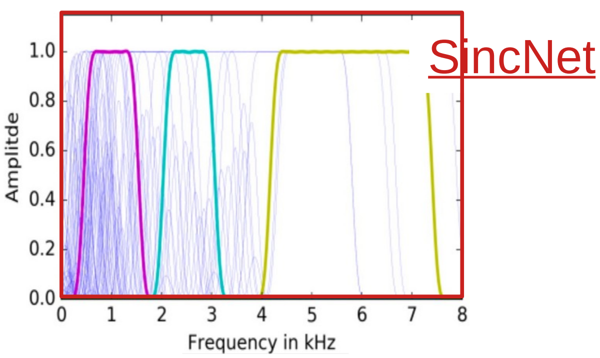
MFCC









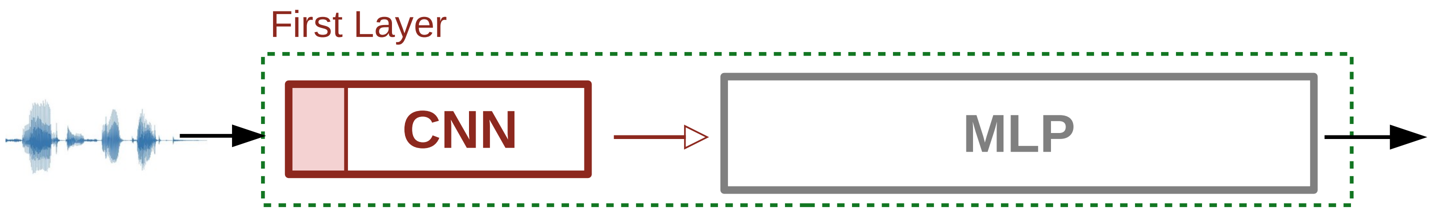


Loweimi et al., et al. On Learning Interpretable CNNs with Parametric Modulated Kernel-based Filters, Interspeech 2019

Listen! 14, Apr, 2020; Parametric CNNs for raw waveform modelling, Slides

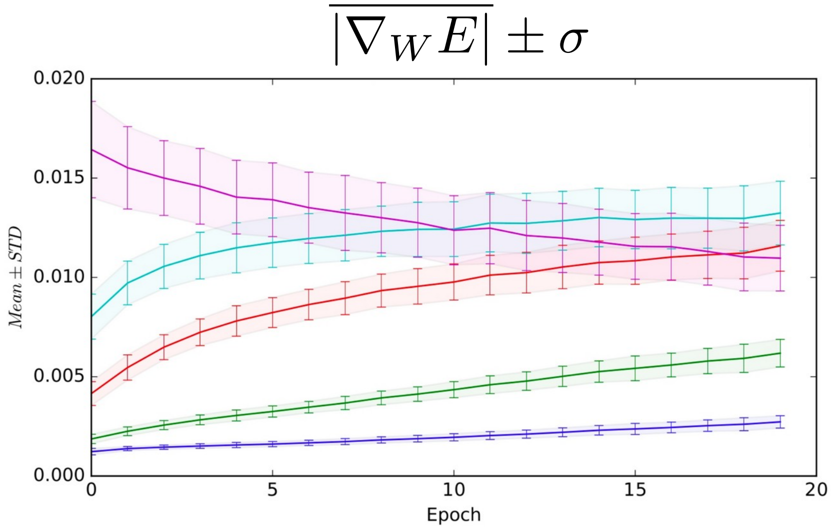
Gradient Vanishing & First Layer

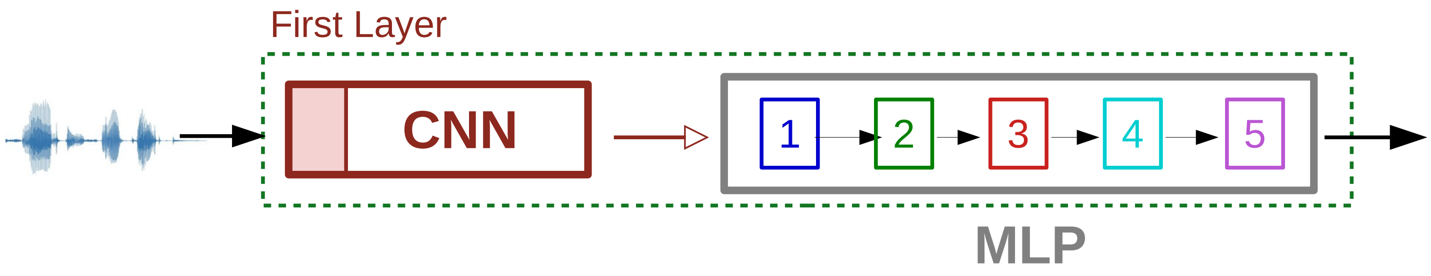
- To what extent is the *gradient vanishing* problematic?



Gradient Vanishing & First Layer

- To what extent is the *gradient vanishing* problematic?





Outline

- Raw waveform acoustic modelling
- Dynamics
  - Dynamics ↔ Temporal evolution ... during training
- Robustness
- Conclusion

First Layer ... TFA ... Questions ...

- To what extent is **it** “vulnerable to gradient vanishing”?

First Layer ... TFA ... Questions ...

- To what extent is **it** “vulnerable to gradient vanishing”?
- What is its training “dynamics” (temporal evolution)?
- How “optimal” are the learned filters?
- How much first layer dynamics correlate with CE/WER?

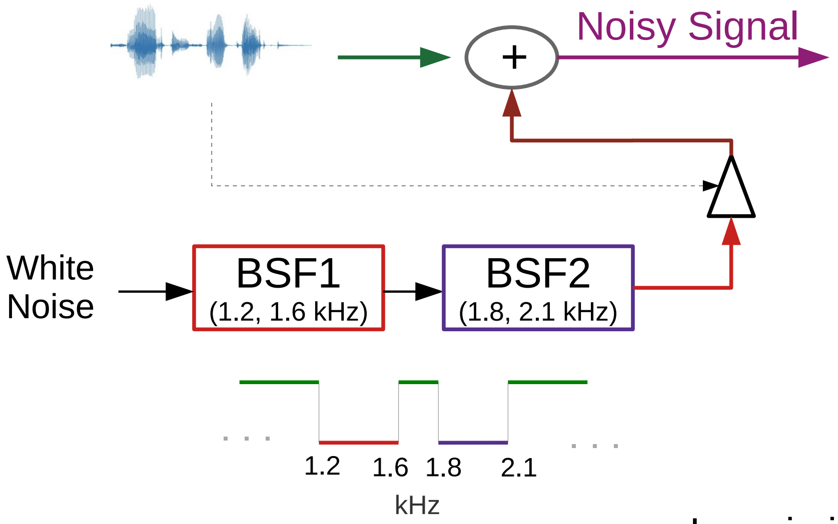
First Layer ... TFA ... Questions ...

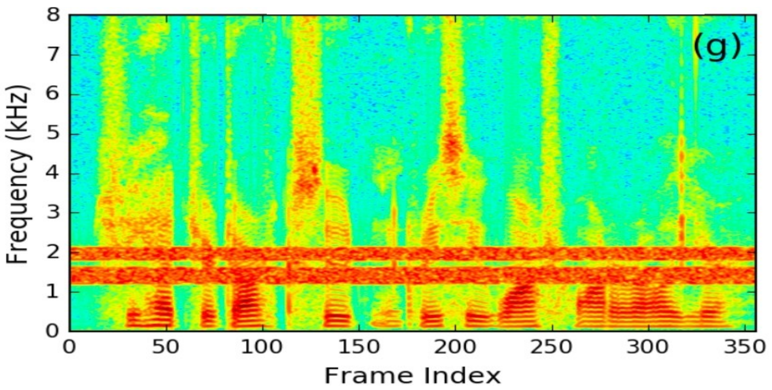
- To what extent is **it** "vulnerable to gradient vanishing"?
- What is its training "dynamics" (temporal evolution)?
- How "optimal" are the learned filters?
- How much first layer dynamics correlate with CE/WER?

- How to investigate all of these?
  - Framework? Task? Metric(s)?

Framework: Task

- Modify TIMIT as follows …
  - Attack two subbands, leave a narrow clean subband in between

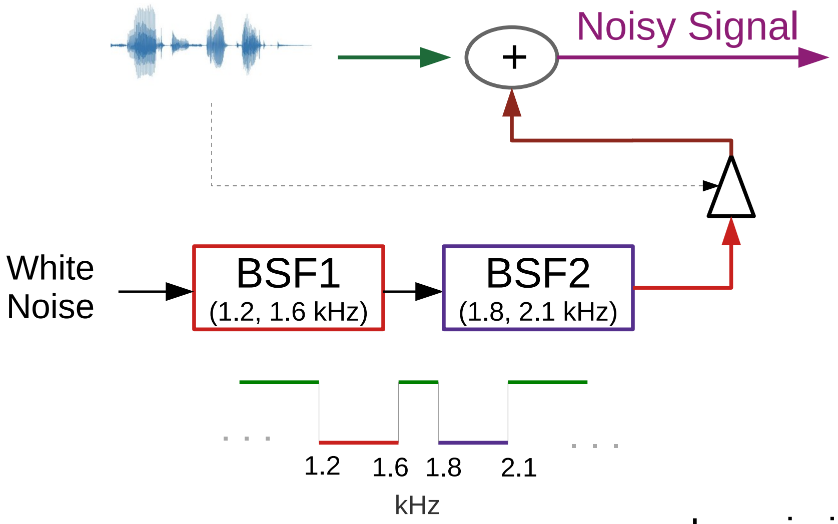


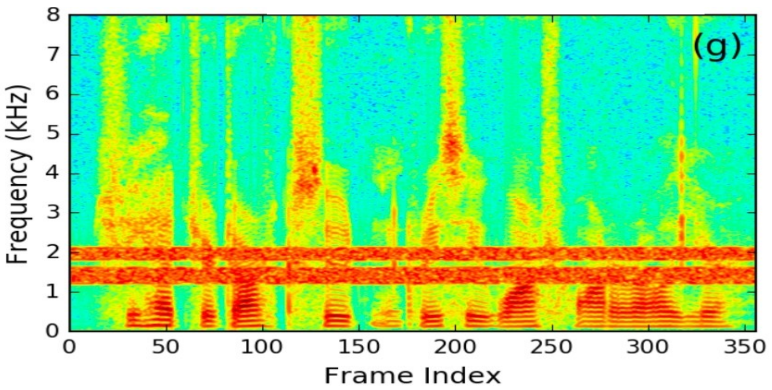


BSF: (ideal) Band Stop Filter

Framework: Task

- Modify TIMIT as follows …
- Advantage: *optimal* solution (TFA) is known



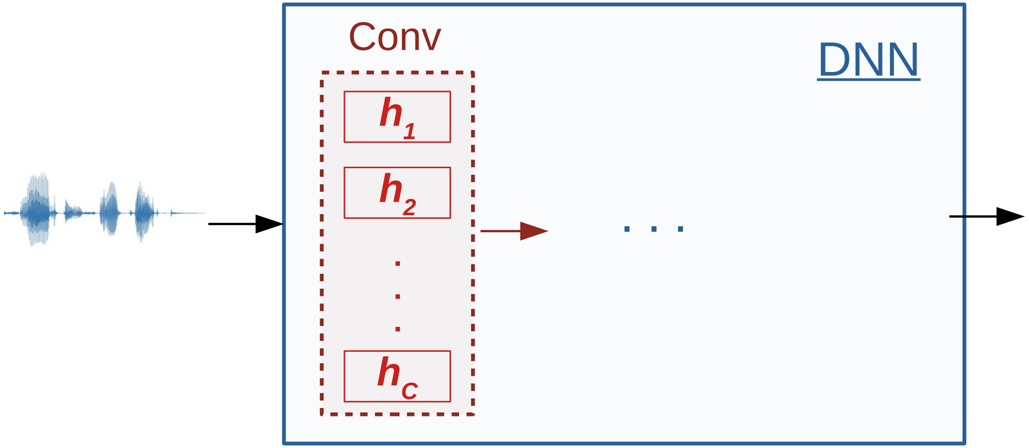


BSF: (ideal) Band Stop Filter

Framework: Metric

- Average Frequency Response (AFR)

$$\mathrm{AFR} = \frac{1}{C}\sum_{c=1}^{C} |H_c(\omega)|$$



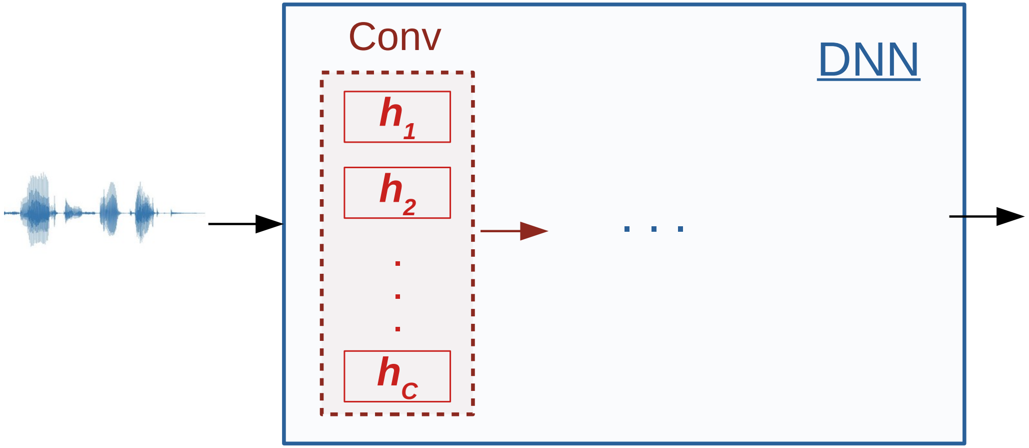
h: impulse response
H: frequency response
C: #channels

Framework: Metric

- Average Frequency Response (AFR)
  - A proxy for the frequency response of the first layer

$$\text{AFR} = \frac{1}{C}\sum_{c=1}^{C} |H_c(\omega)|$$

h: impulse response
H: frequency response
C: #channels

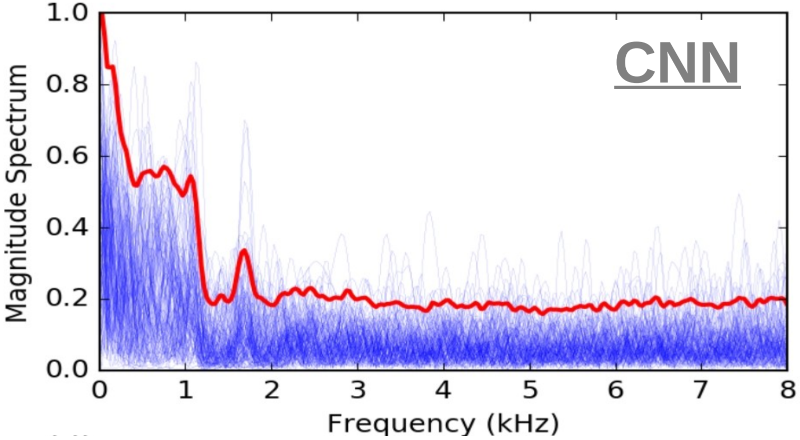


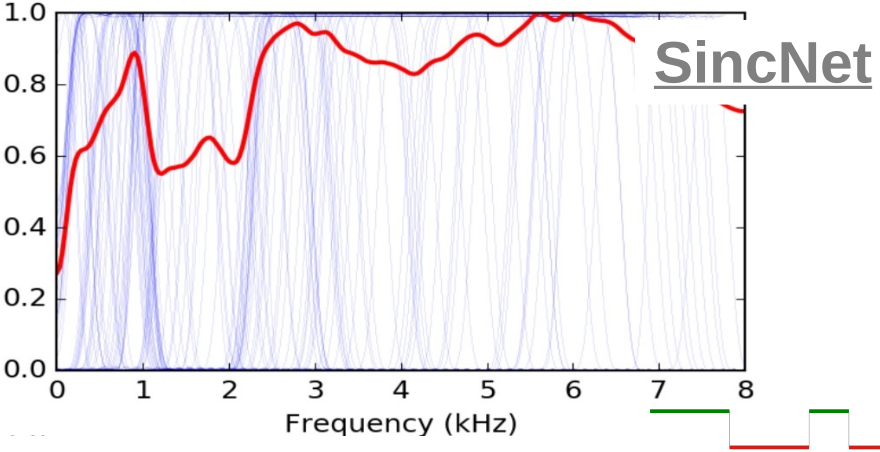
Setup

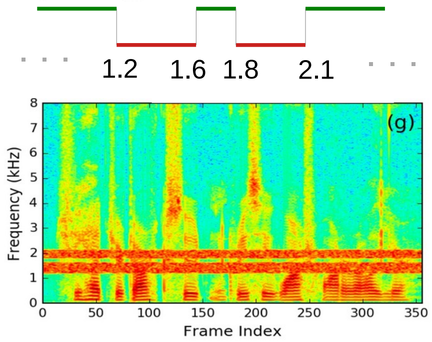
- Raw waveform models: CNN and SincNet
- Database: TIMIT, Aurora-4 and WSJ
- Noise: AWGN* → BSF†1 → BSF†2 → SNR: 0 dB
- DNN: CNN-1D (4L) → FC (5L) → Softmax
- Toolkit: PyTorch-Kaldi, default setting

AWGN*: Additive White Gaussian Noise
BSF†: (ideal) Band Stop Filter

AFR … 1^{st} epoch

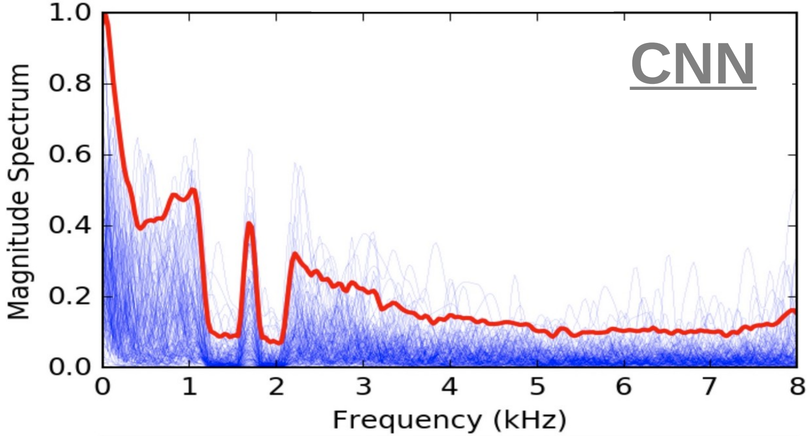


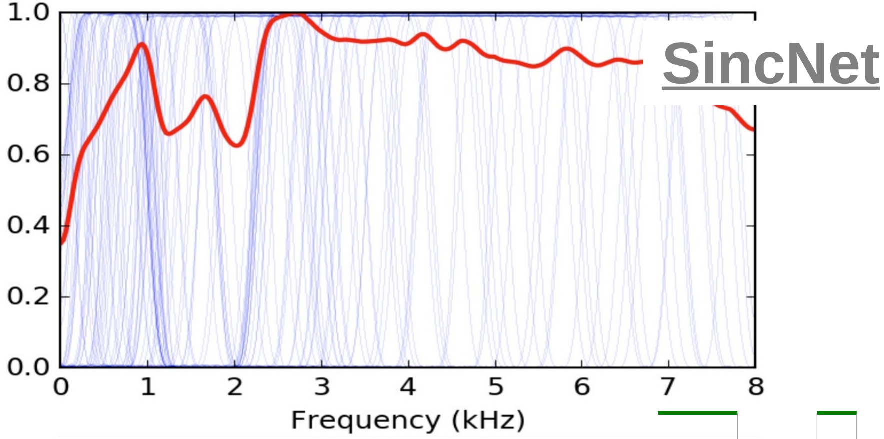


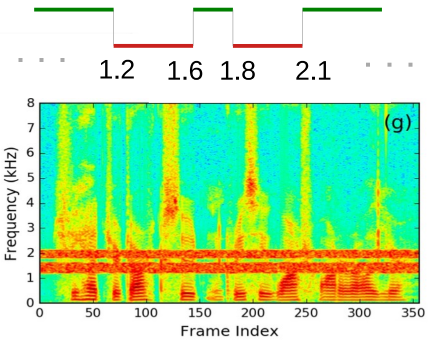


- SincNet approx. finds the noisy subbands
  - Learns faster than CNN ← fewer params

AFR … 20th epoch

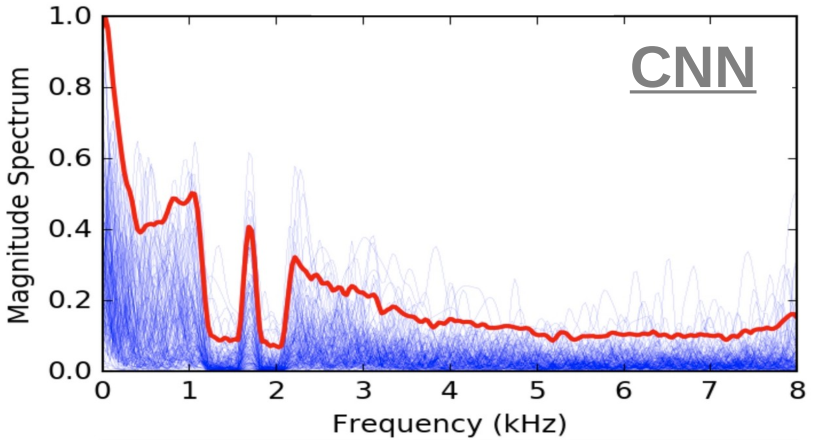


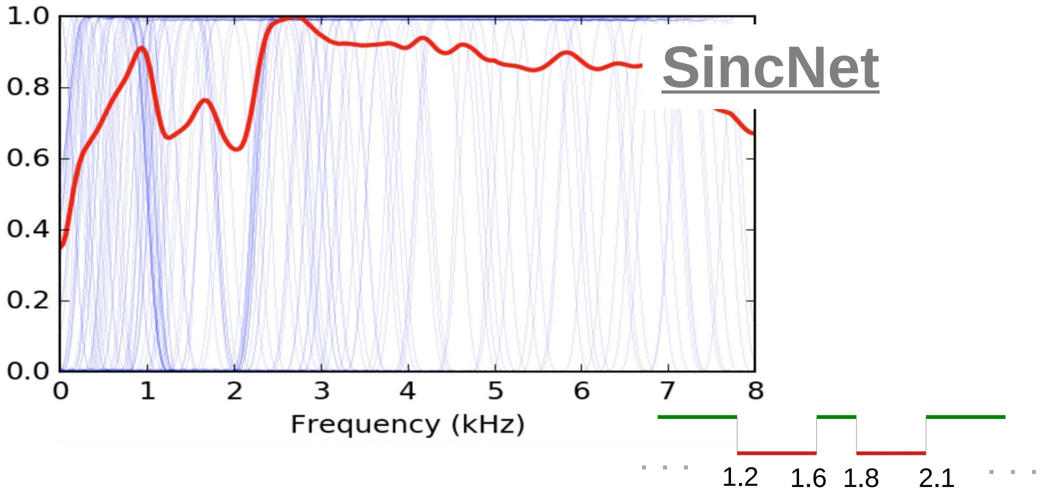




- Both find out the noisy and clean subbands
- CNN has a higher spectral resolution

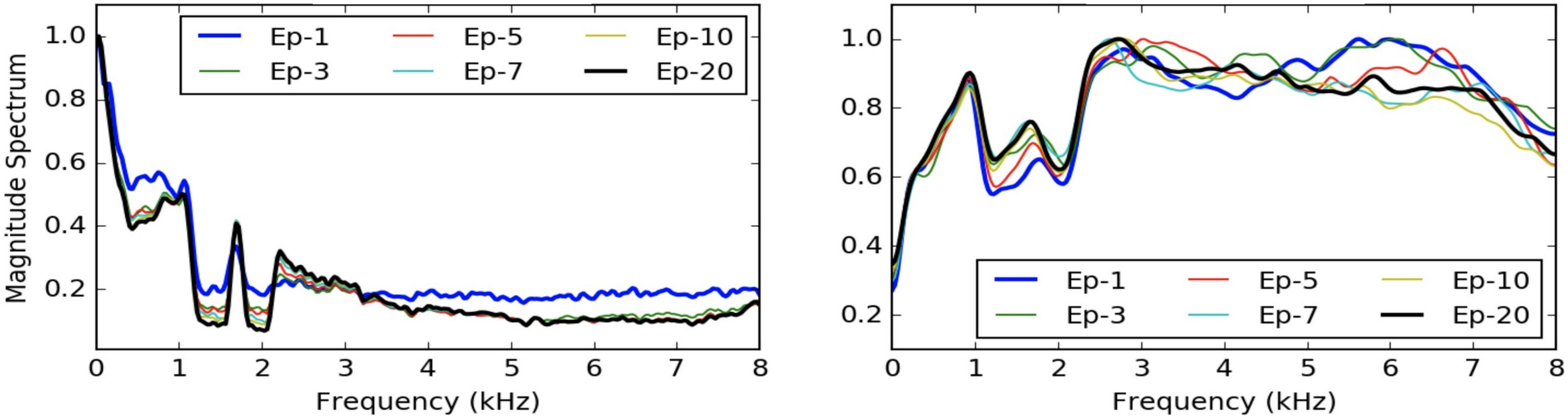
AFR … 20th epoch



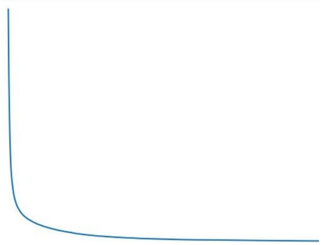


- Both find out the noisy and clean subbands
- Solving an enhancement problem using ASR labels (?)

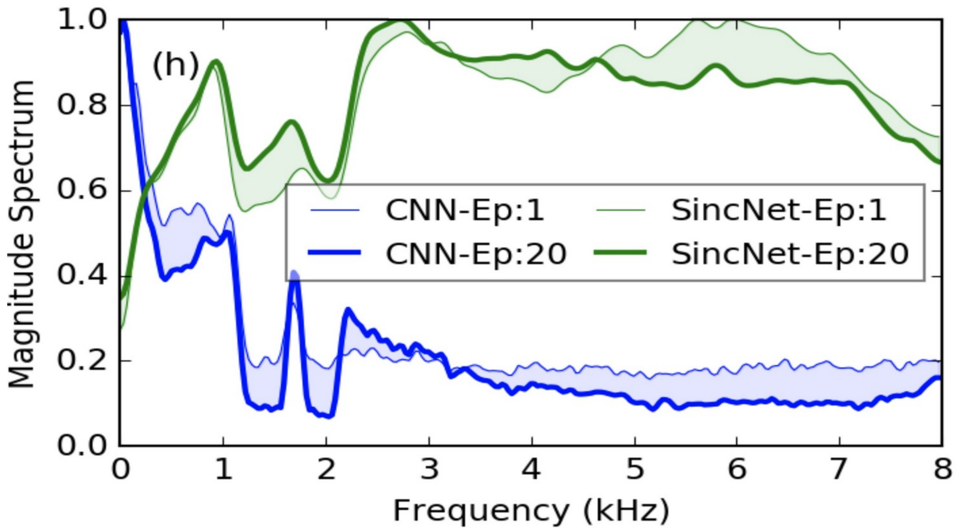
Temporal Evolution of AFR (1)

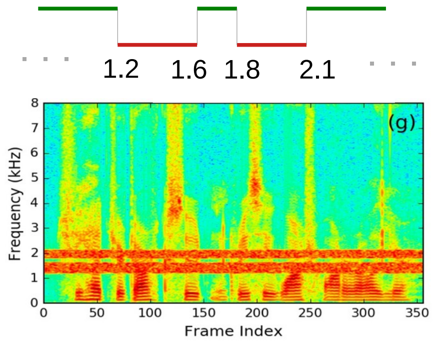


- AFR change rate reduced for higher epochs
- After 10 epochs, AFR converges



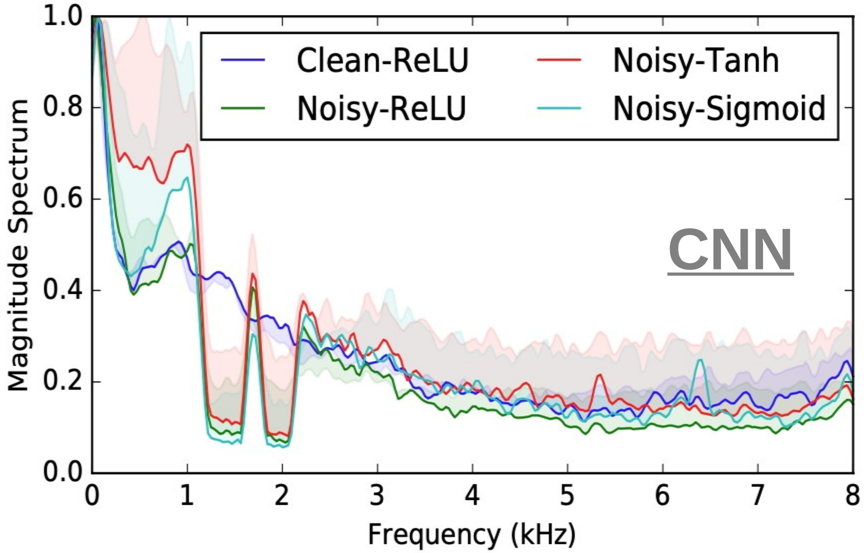
Temporal Evolution of AFR (2)

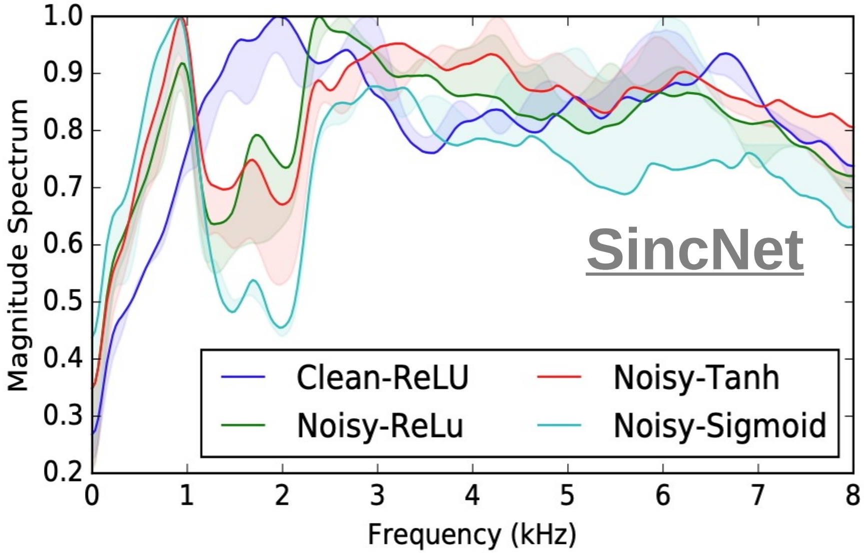




Shaded area between epoch 1 to 20 ≡ Training Dynamics

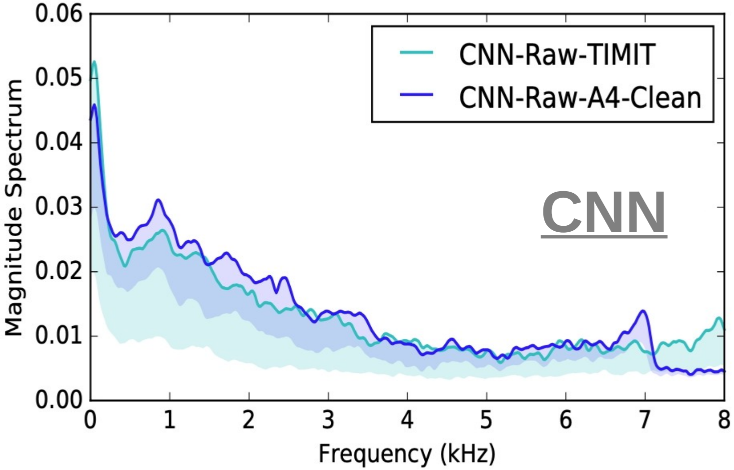
Effect of Non-linearity

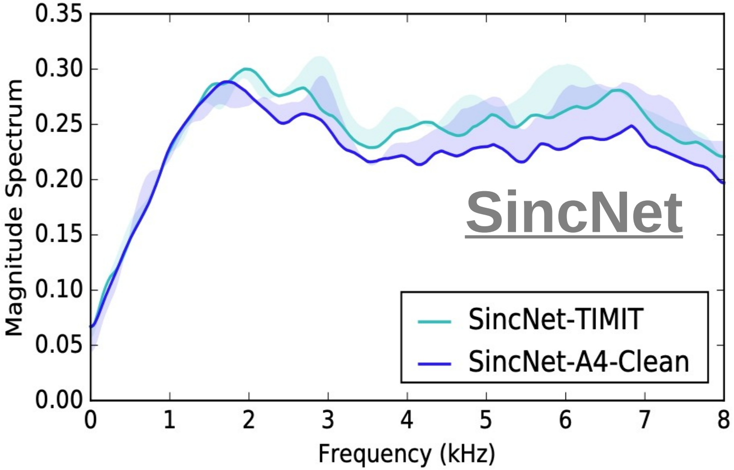




- Tanh & Sigmoid → larger shaded area → slower convergence
- ReLU → smaller shaded area (CNN) → faster conv ← Sparsity

Database Effect: TIMIT vs Aurora-4 (A4)

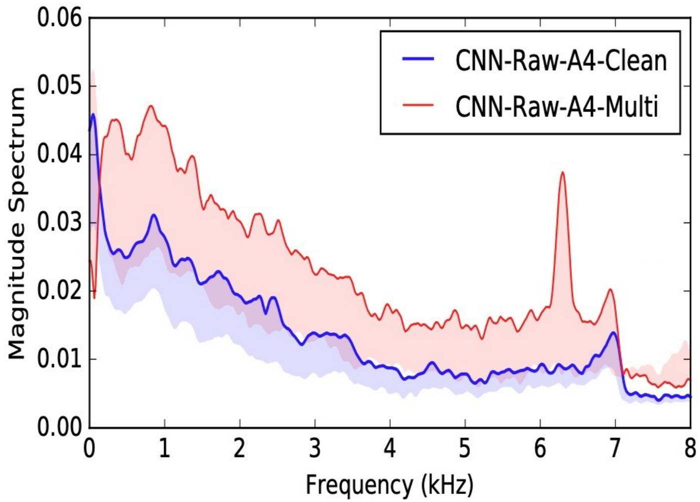




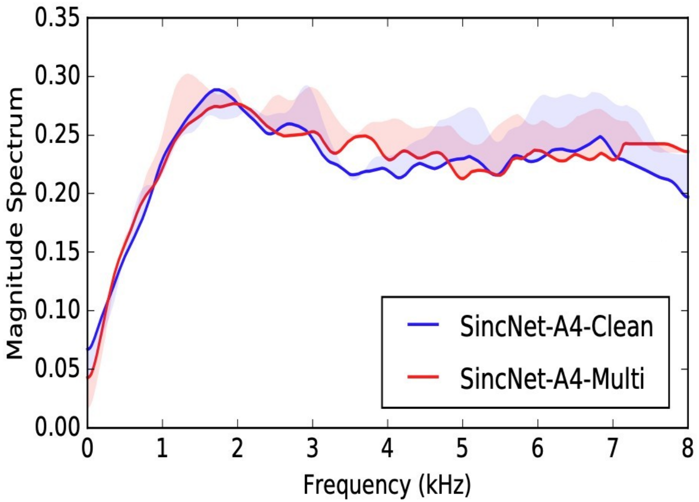
- AFR for A4-Clean and TIMIT are almost similar
- Shaded area for A4 is smaller, especially for CNN-Raw

Database Effect: A4, Clean vs Multi

CNN

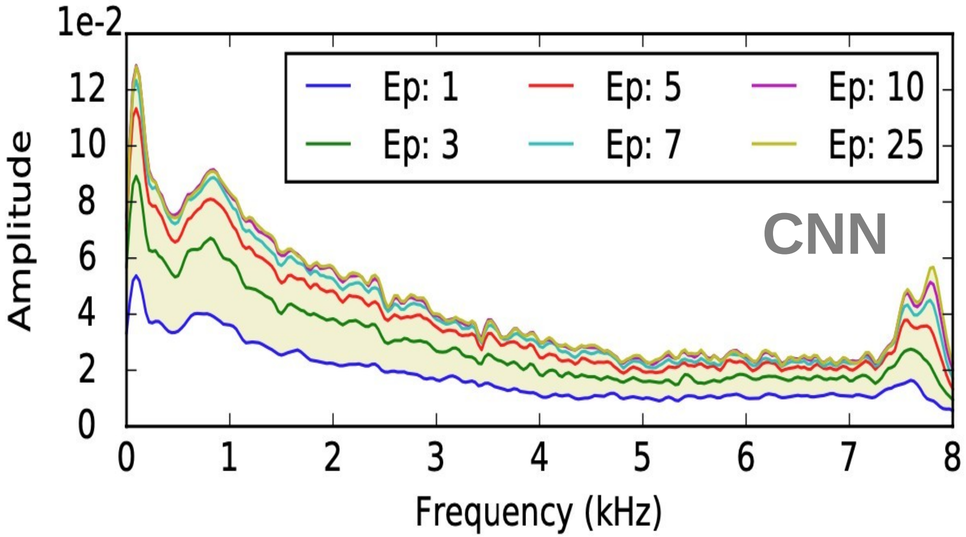


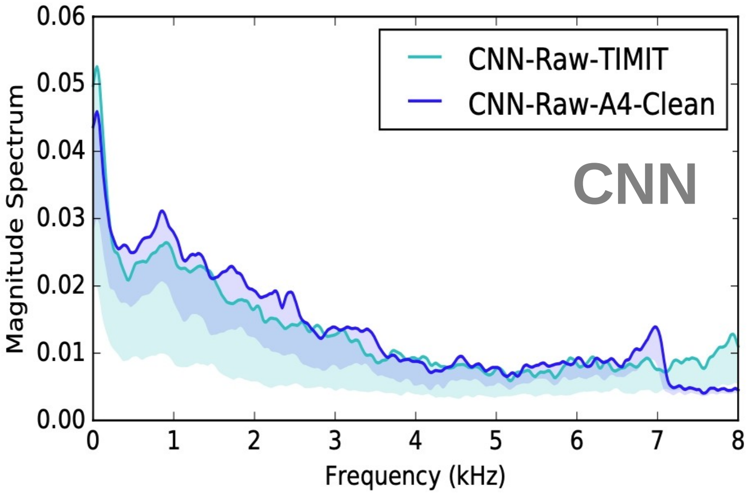
SincNet



- Shaded area is larger for A4 Multi-style
  - Richer variability → More to learn!

Database Effect: WSJ





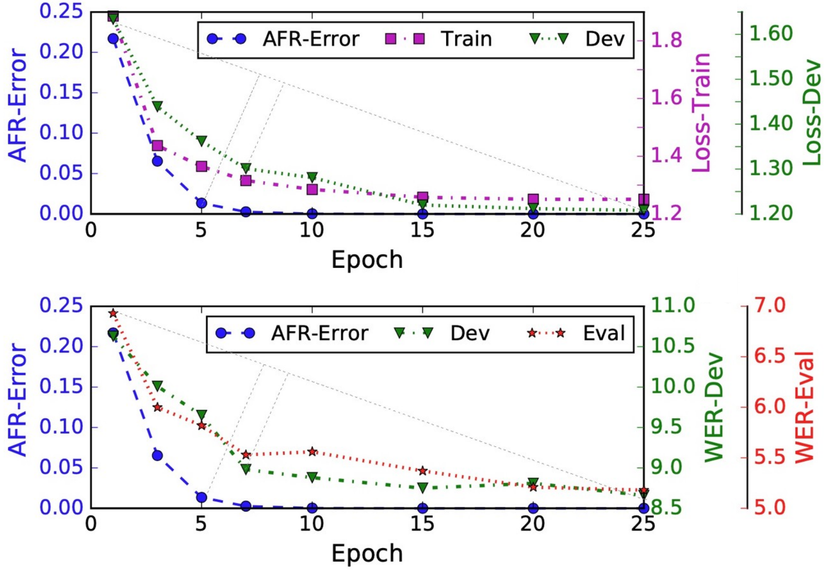
- AFR is almost similar for these databases (all clean)

Correlation of AFR & {CE,WER}

- Database: WSJ
- $AFR_{Error} = MSE\{AFR_{ep} - AFR_{optimal}\}$
  - Assuming $AFR_{optimal} \equiv AFR_{25}$

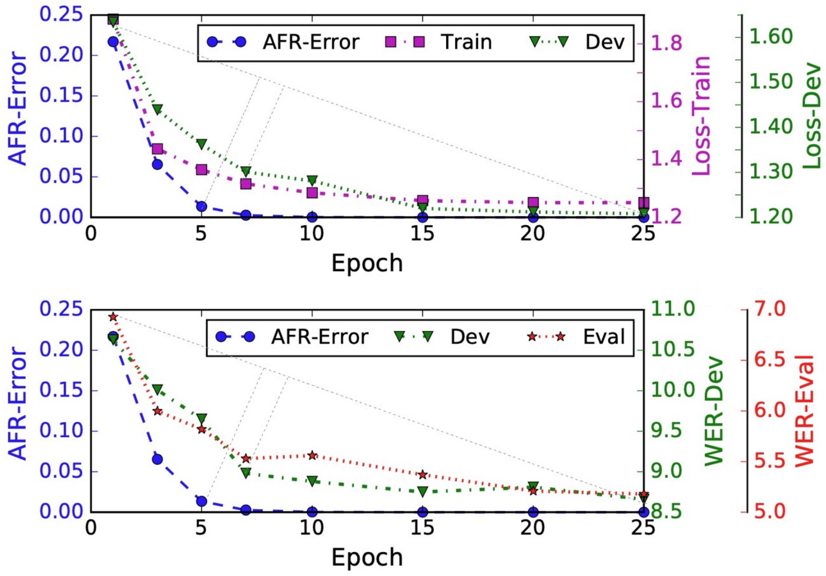
Correlation of AFR & {CE,WER}

- Database: WSJ
- $AFR_{Error} = MSE\{AFR_{ep} - AFR_{25}\}$
- Similar dynamics … knee points ...



Correlation of AFR & {CE,WER}

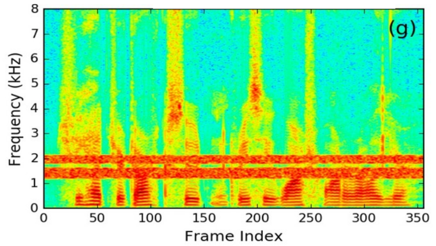
- Database: WSJ
- $AFR_{Error} = MSE\{AFR_{ep} - AFR_{25}\}$
- Similar dynamics … knee points ...
- AFR temporal evolution highly correlates with CE/WER dynamics

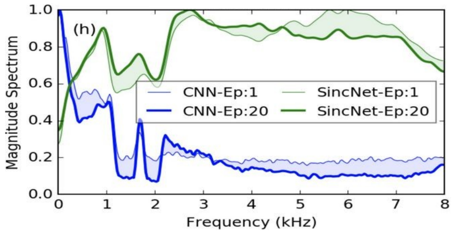


	CE-Train	CE-Dev	WER-Dev	WER-Eval
Corr	0.99	0.94	0.88	0.95

Outline

- Raw waveform acoustic modelling
- Dynamics
- Robustness
  - How robust the raw waveform models are?
  - How the performance can be improved?
- Conclusion



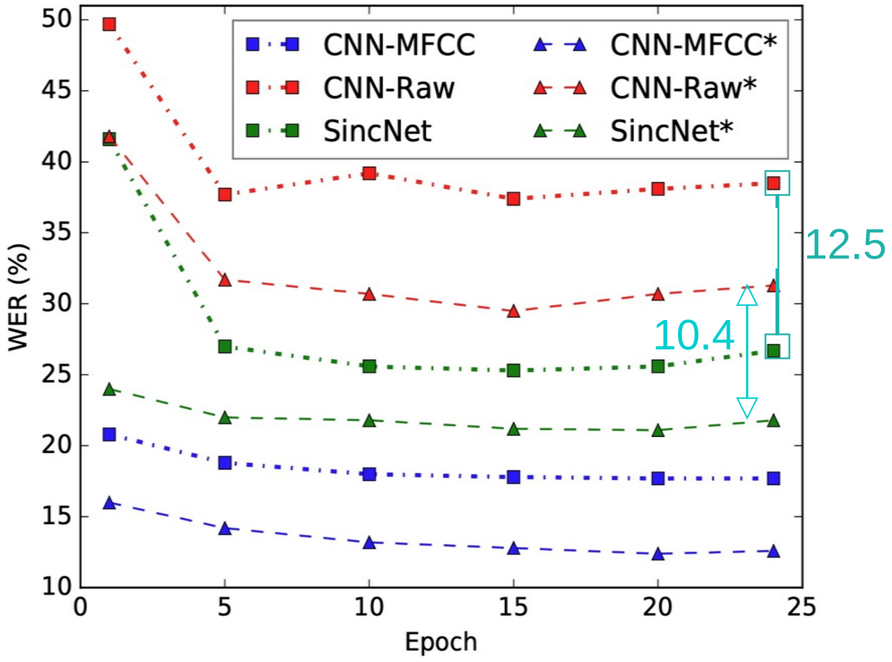


Setup

- DNNs built using PyTorch-Kaldi
- Databases: TIMIT, Aurora-4, WSJ
- Frame length/shift: 25/10ms $\leftrightarrow$ MFCC; 200/10ms $\leftrightarrow$ Raw wave
- Context length: ±5 for MFCC, 0 for raw waveform
- Feature normalisation for raw waveform was done dimension-wise, similar to MFCC
  - * → Mean-Var Normalisation at utterance level
  - † → Mean-Var Normalisation at speaker level

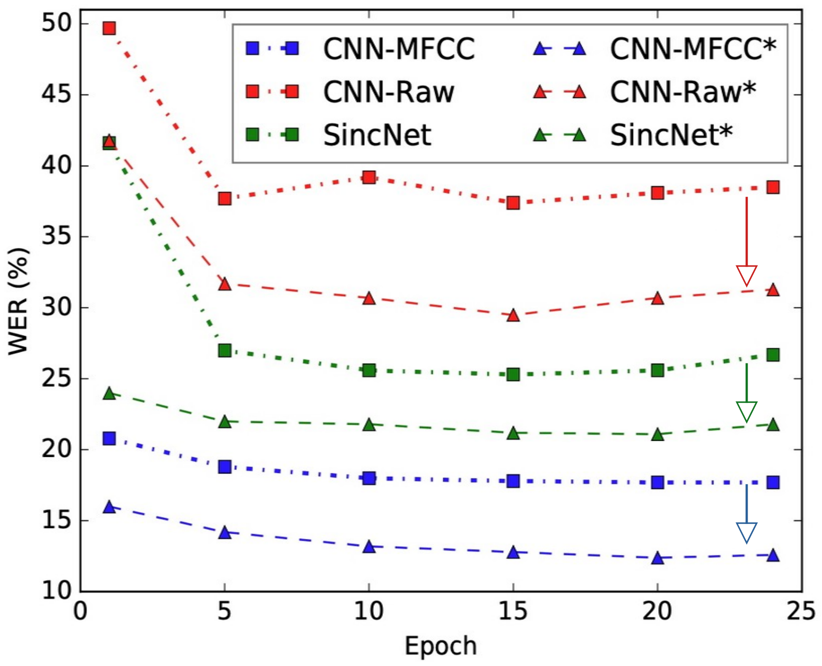
Aurora-4, Clean Training

- $\text{WER}_{\text{MFCC}} < \text{WER}_{\text{FBank}} < \text{WER}_{\text{Raw}}$
- WER **gap** between SincNet and CNN-raw is large



Aurora-4, Clean Training

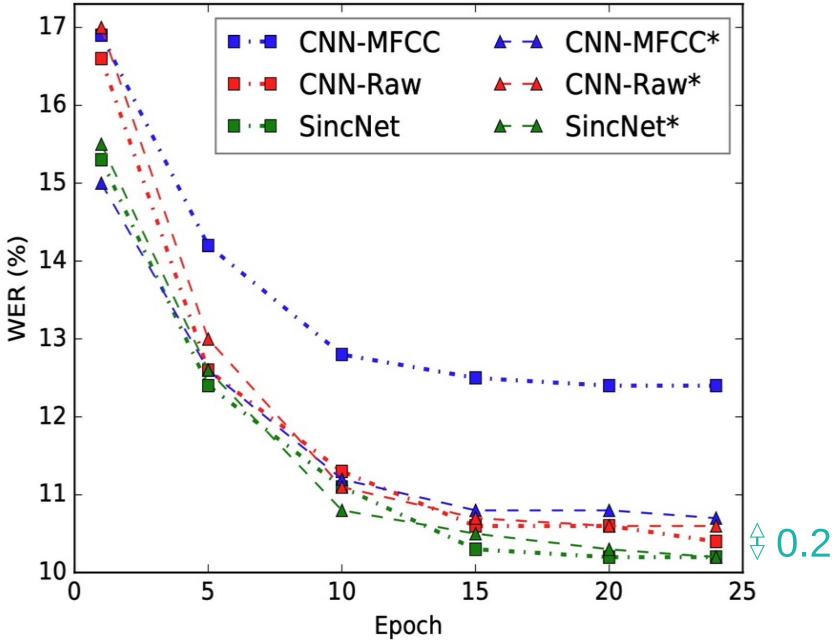
- $WER_{MFCC} < WER_{FBank} < WER_{Raw}$
- WER **gap** between SincNet and CNN-raw is large
- MVN* helpful for all ...
  - [abs, Rel.] Gain in % (epoch 25)
    - MFCC → [5.1, 30.0]
    - CNN → [7.5, 19.4]
    - SincNet → [4.3, 16.8]



MVN*: mean-var norm at utter level

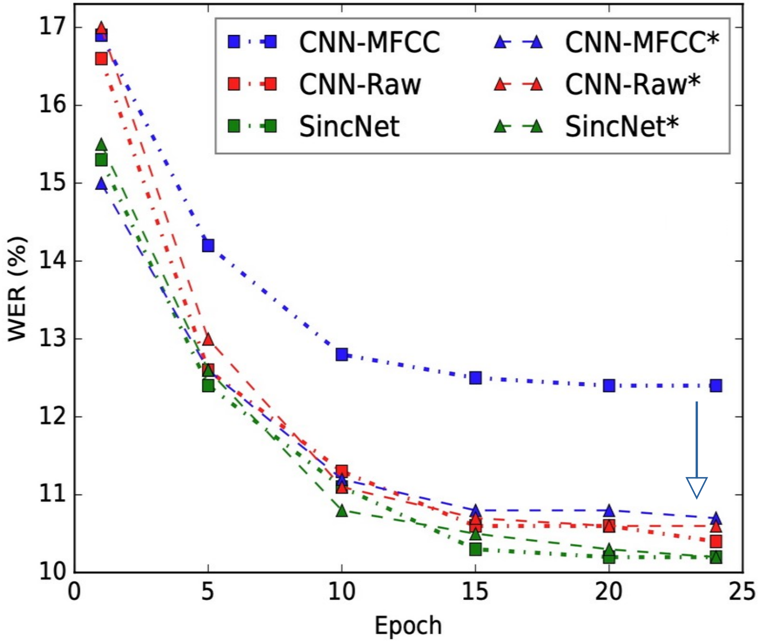
Aurora-4, Multi-condition Training

- $WER_{FBank} < WER_{Raw} < WER_{MFCC}$
- WER **gap** between CNN and SincNet is very small



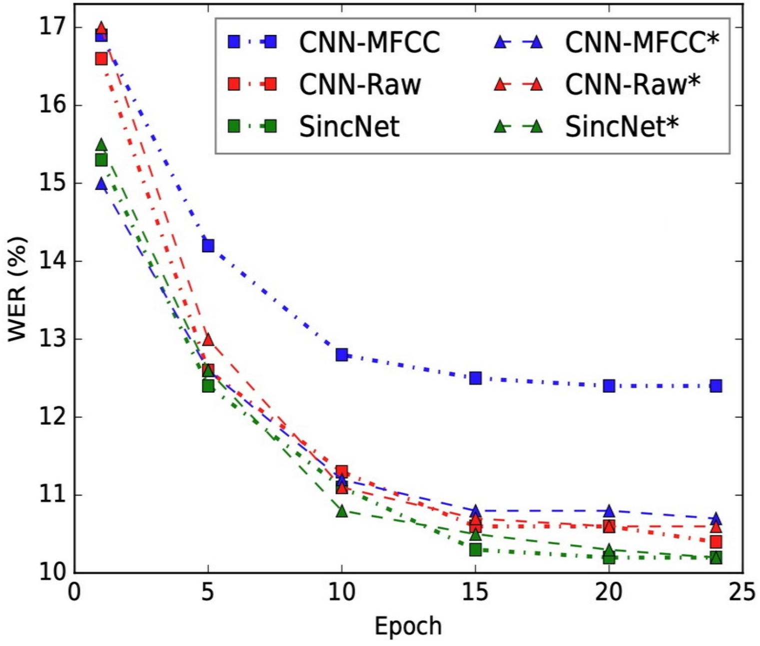
Aurora-4, Multi-condition Training

- $WER_{FBank} < WER_{Raw} < WER_{MFCC}$
- WER **gap** between CNN and SincNet is very small
- Feature normalisation ...
  - helpful for MFCC
  - does **NOT** help raw waveform



Aurora-4, Multi-condition Training

- $WER_{FBank} < WER_{Raw} < WER_{MFCC}$
- WER **gap** between CNN and SincNet is very small
- Feature normalisation ...
  - helpful for MFCC
  - does **NOT** help raw waveform
- How can we reduce WER?



A Detour → WSJ

- **Detour** → WSJ is not for robustness!
- Raw waveform outperforms others
  - $\mathrm{WER}_{\mathrm{Raw}} < \mathrm{WER}_{\mathrm{FBank}} < \mathrm{WER}_{\mathrm{MFCC}}$

Table 2: *WSJ WER for different front-ends.*

	MFCC†	FBank†	CNN-Raw	Sinc-Raw
Dev93	10.4	9.1	8.6	8.5
Eval92	6.8	5.9	5.1	5.0

A Detour → WSJ

- **Detour** → WSJ is not for robustness
- Raw waveform outperforms others
  - $\text{WER}_{\text{Raw}} < \text{WER}_{\text{FBank}} < \text{WER}_{\text{MFCC}}$
- **Why**? More data (81 h)

Table 2: *WSJ WER for different front-ends.*

	MFCC†	FBank†	CNN-Raw	Sinc-Raw
Dev93	10.4	9.1	8.6	8.5
Eval92	6.8	5.9	5.1	5.0

A Detour → WSJ

- **Detour** → WSJ is not for robustness
- Raw waveform outperforms others
  - $\text{WER}_{\text{Raw}} < \text{WER}_{\text{FBank}} < \text{WER}_{\text{MFCC}}$
- **Why**? More data (81 h)
  - **ONLY** data amount? TIMIT → …

Table 2: *WSJ WER for different front-ends.*

	MFCC†	FBank†	CNN-Raw	Sinc-Raw
Dev93	10.4	9.1	8.6	8.5
Eval92	6.8	5.9	5.1	5.0

Table 2: *TIMIT PER for different kernels (200ms).*

	MLP	CNN	Sinc	Sinc^2	Gamma	Gauss
PER	18.5	18.2	17.6	16.9	17.2	17.0

Loweimi et al., et al. On Learning Interpretable CNNs with Parametric Modulated Kernel-based Filters, Interspeech 2019

A Detour → WSJ

- **Detour** → WSJ is not for robustness
- Raw waveform outperforms others
  - $\text{WER}_{\text{Raw}} < \text{WER}_{\text{FBank}} < \text{WER}_{\text{MFCC}}$
- **Why**? More data (81 h)
  - **ONLY** data amount? TIMIT → …
- **Hypothesis**:
  - Teacher/label error is more problematic for high-dim features

Table 2: *WSJ WER for different front-ends.*

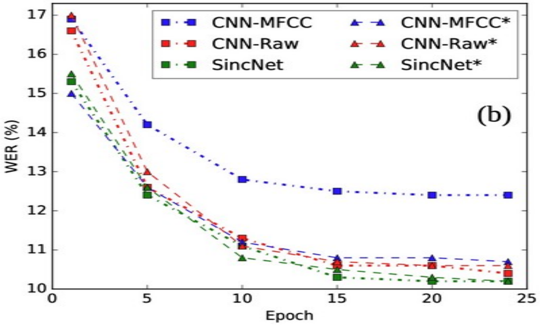
	MFCC†	FBank†	CNN-Raw	Sinc-Raw
Dev93	10.4	9.1	8.6	8.5
Eval92	6.8	5.9	5.1	5.0

Table 2: *TIMIT PER for different kernels (200ms).*

	MLP	CNN	Sinc	Sinc^2	Gamma	Gauss
PER	18.5	18.2	17.6	16.9	17.2	17.0

Back to Aurora-4, Multi-condition

Alignment from Multi



$WER_{FBank} < WER_{Raw} < WER_{MFCC}$

- Reduce teacher/label error via using a better alignment
- Better alignment obtained using clean training data

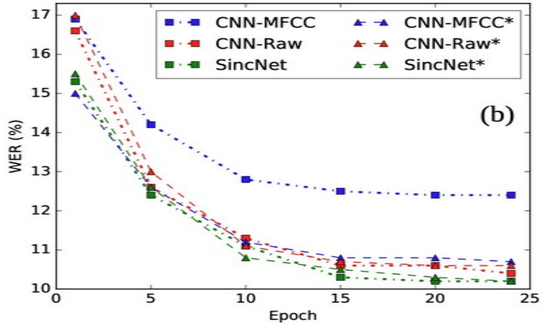
Back to Aurora-4, Multi-condition

Alignment from Clean

Feature	A	B	C	D	*Ave*
CNN-MFCC*	3.5	6.1	4.6	8.3	6.7
CNN-FBank*	3.0	5.2	3.3	6.4	5.4
CNN-Raw	2.7	4.4	4.0	6.4	5.1
SincNet-Raw	2.9	4.6	3.9	6.7	5.3

$WER_{Raw} < WER_{FBank} < WER_{MFCC}$

Alignment from Multi



$WER_{FBank} < WER_{Raw} < WER_{MFCC}$

- Reduce teacher/label error via using a better alignment
- Better alignment obtained using clean training data

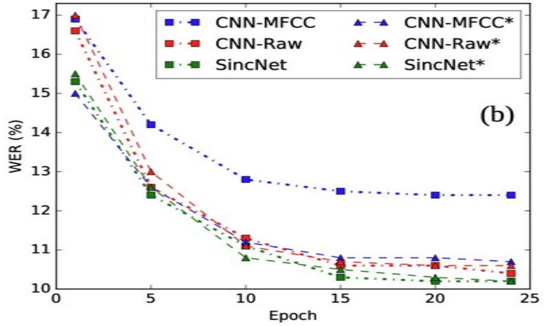
Back to Aurora-4, Multi-condition

Alignment from Clean

Feature	A	B	C	D	*Ave*
CNN-MFCC*	3.5	6.1	4.6	8.3	6.7
CNN-FBank*	3.0	5.2	3.3	6.4	5.4
CNN-Raw	2.7	4.4	4.0	6.4	5.1
SincNet-Raw	2.9	4.6	3.9	6.7	5.3

$WER_{Raw} < WER_{FBank} < WER_{MFCC}$

Alignment from Multi



$WER_{FBank} < WER_{Raw} < WER_{MFCC}$

- Reduce teacher/label error via using a better alignment
- Better alignment obtained using clean training data ...

  … is more beneficial to raw waveform models

Outline

- Raw waveform acoustic modelling for ASR
- Dynamics
- Robustness
- Conclusion

Conclusion

- **Keywords**: ASR, Raw waveform, Dynamics, Robustness
- **Dynamics** ≡ Temporal evolution ... first conv layer
  - Task: TIMIT+ Special Noise
  - Metric: Average Frequency Response (AFR)
  - What was studied: Gradient vanishing, optimality, resolution, non-linearity, database, correlation of AFR with CE & WER
- **Robustness**
  - Mismatched condition → feature normalisation
  - Matched condition → better alignment (lower teacher error)

That’s It!

- Thanks for your attention!
- Q/A?

- Paper link

SpeechWave