



# Speech Acoustic Modelling from Raw Signal Representations

Erfan Loweimi

CogM Hear Workshop, Edinburgh Napier University  
19, Oct, 2022

# SpeechWave



Loweimi et al

# Outline

- Motivation
- Architecture
- Variants, Analysis & Interpretation
- Conclusion

# Outline

- Motivation
- Architecture
- Variants, Analysis & Interpretation
- Conclusion

# Motivation ...

# Perfect Information Processing ...

# (1) Perfect Info Filtering ...

- Pass **signal**, Discard **noise**

\* Signal: task-correlated info  
\* Noise: task-irrelevant info

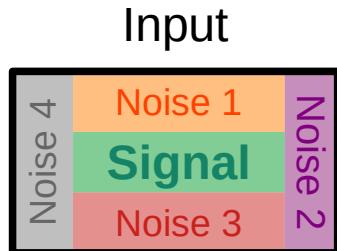
# (1) Perfect Info Filtering ...

- Pass **signal**, Discard **noise**
  - **Discriminability, Robustness/Generalisation**

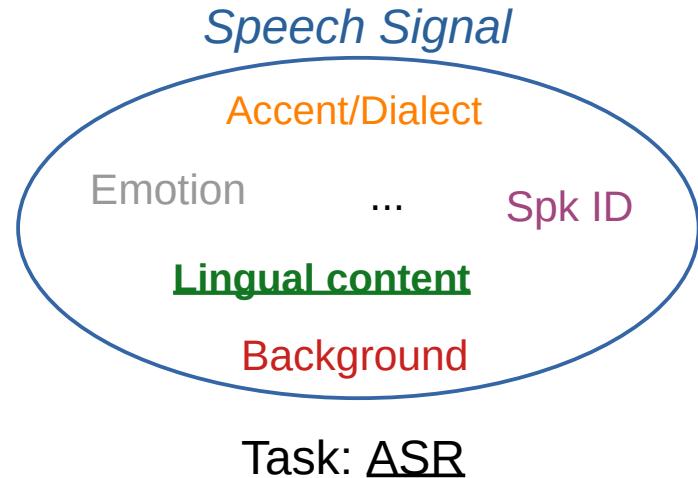
\* Signal: task-correlated info  
\* Noise: task-irrelevant info

# (1) Perfect Info Filtering ...

- Pass **signal**, Discard **noise**
  - **Discriminability, Robustness/Generalisation**

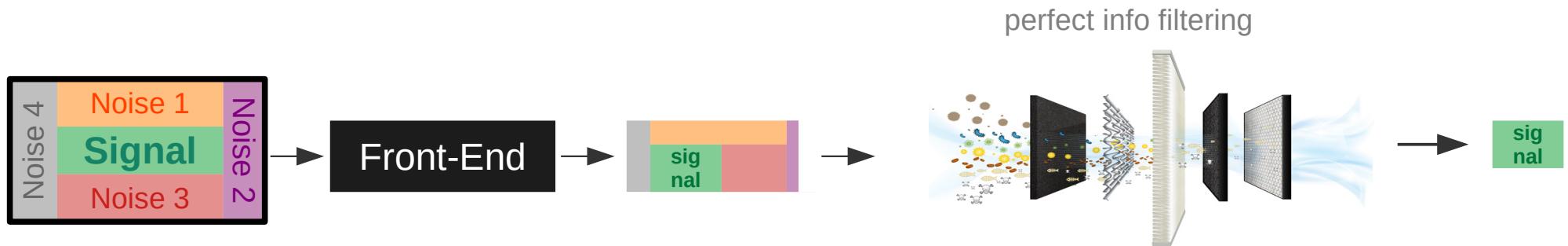


\* Signal: task-correlated info  
\* Noise: task-irrelevant info



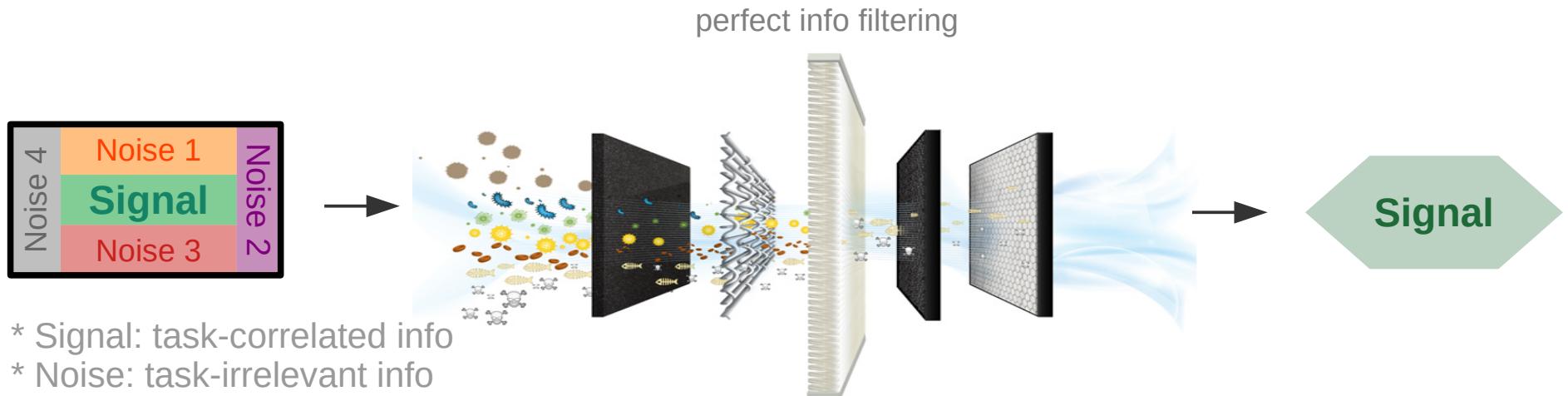
# (3) Perfect Input ...

- Garbage in, Garbage out ...
  - “... output can only be as accurate as the info entered ... ”



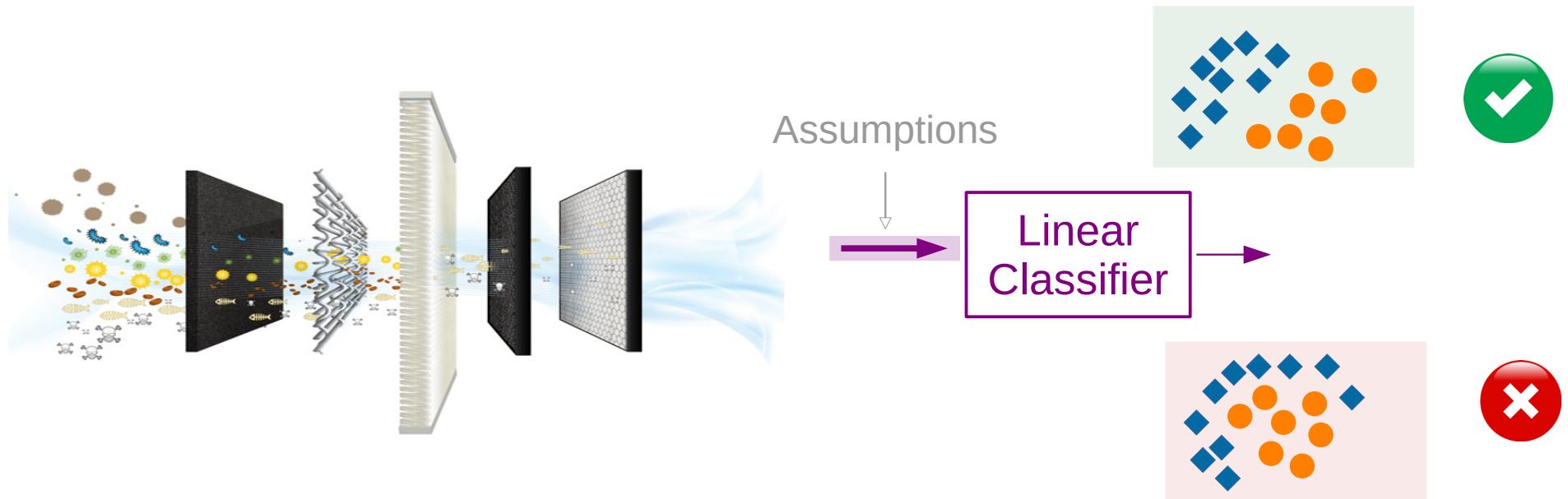
# (1) Perfect Info Filtering ...

- Pass **signal**, Discard **noise**
  - **Discriminability, Robustness/Generalisation**

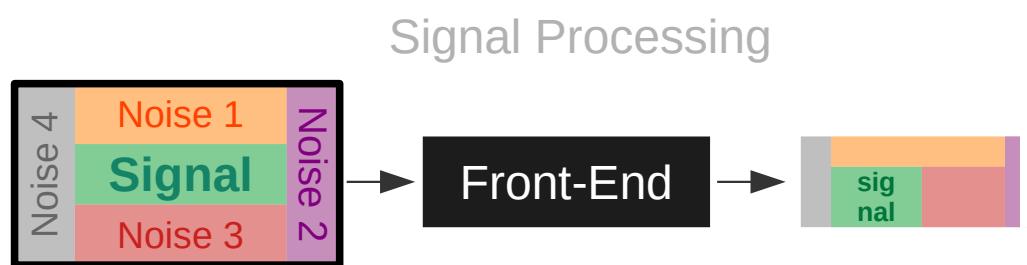


## (2) Perfect Match with Classifier ...

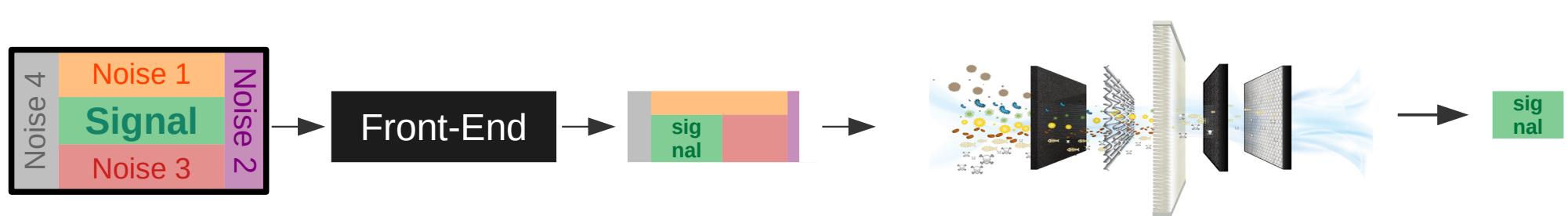
- SoftMax  $\leftrightarrow$  Linear classifier  $\leftarrow$  Linear separability



# (3) Perfect Input ...

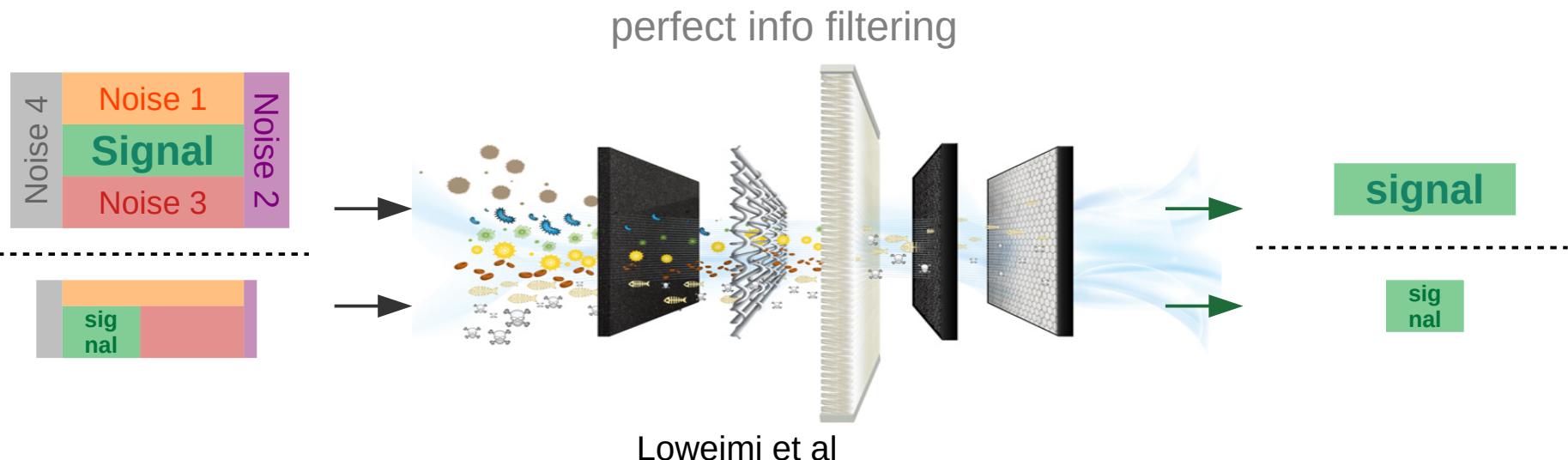


# (3) Perfect Input ...



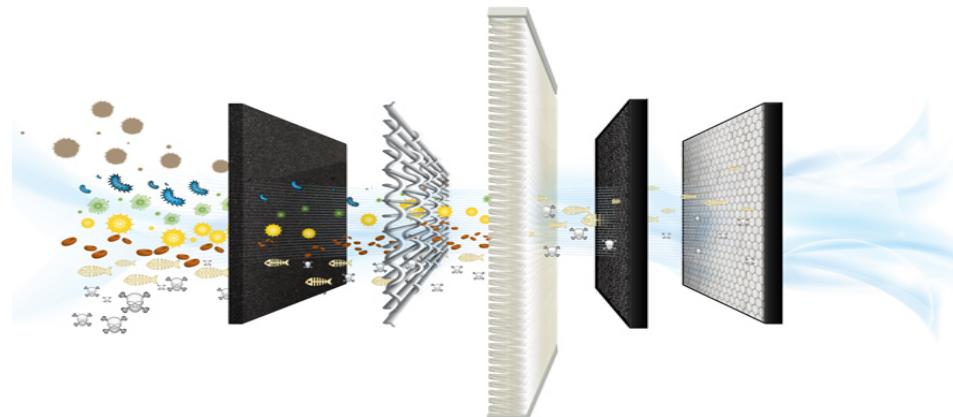
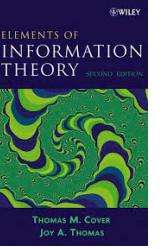
# (3) Perfect Input ...

- Garbage in, Garbage out ...
  - “... output can only be as accurate as the info entered ... ”



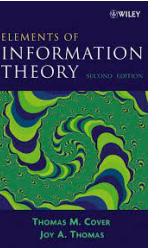
# (3) Perfect Input ...

- Data Processing Inequality (DPI)
  - “... *Processing cannot increase information ...*”

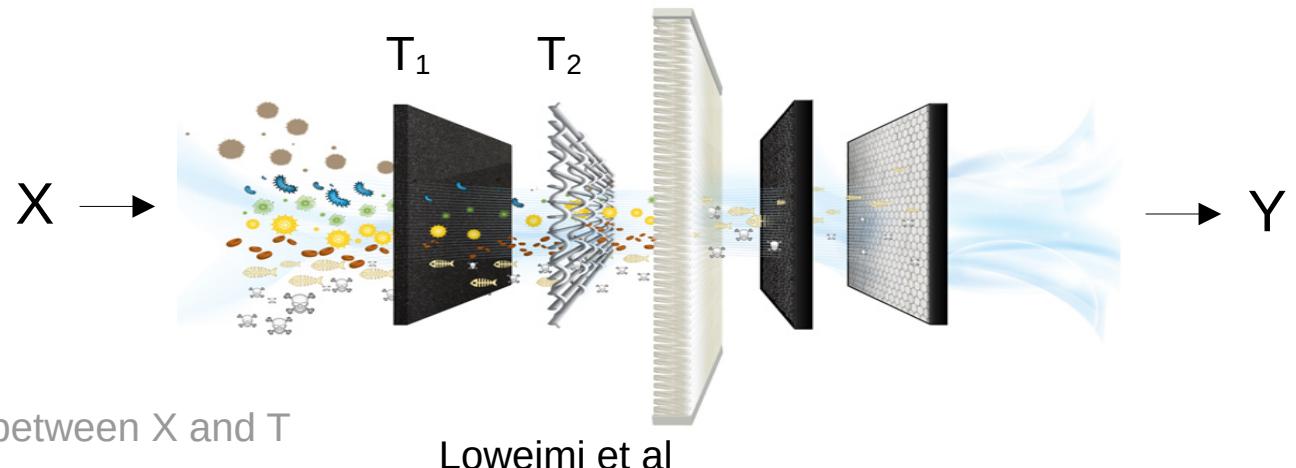


Loweimi et al

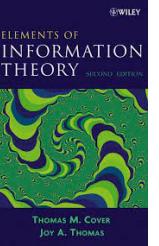
# (3) Perfect Input ...



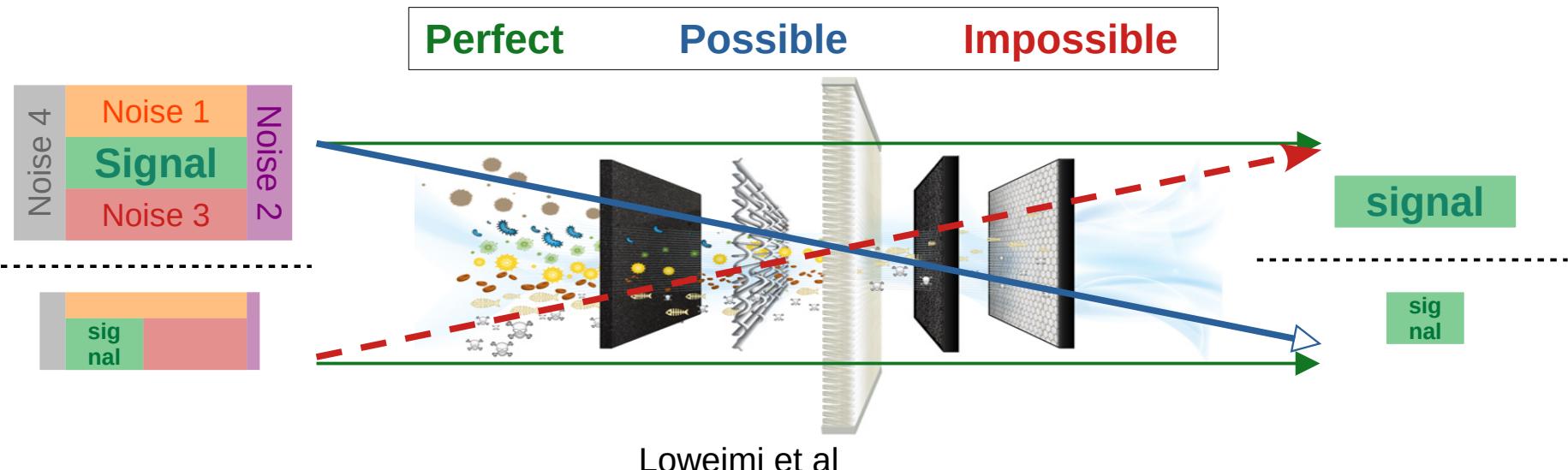
- Data Processing Inequality (DPI)
  - “... *Processing cannot increase information* ...”
  - Markov Chain:  $X \rightarrow T_1 \rightarrow T_2 \rightarrow \dots \implies I(X;T_1) \geq I(X;T_2) \geq \dots$



# (3) Perfect Input ...



- Data Processing Inequality (DPI)
  - “... Processing cannot increase information ...”
  - Markov Chain:  $X \rightarrow T_1 \rightarrow T_2 \rightarrow \dots \implies I(X;T_1) \geq I(X;T_2) \geq \dots$



# Building A Perfect System Requires ...

- Perfect Filtering
- Perfect Match
- Perfect Input

# Building A Perfect System Requires ...

- Perfect Filtering  $\leftrightarrow$  Architecture+Data+Training
- Perfect Match  $\leftrightarrow$  Architecture+Data+Training
- Perfect Input  $\leftarrow$  include task-useful ...

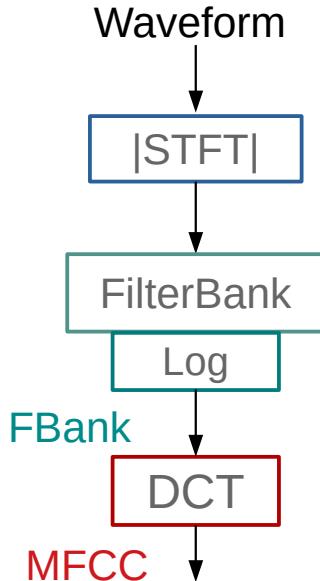
# Building A Perfect System Requires ...

- Perfect Filtering  $\leftrightarrow$  Architecture+Data+Training
- Perfect Match  $\leftrightarrow$  Architecture+Data+Training
- Perfect Input  $\leftarrow$  includes task-useful; possible?

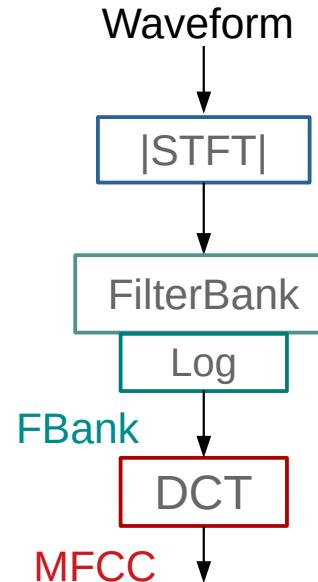
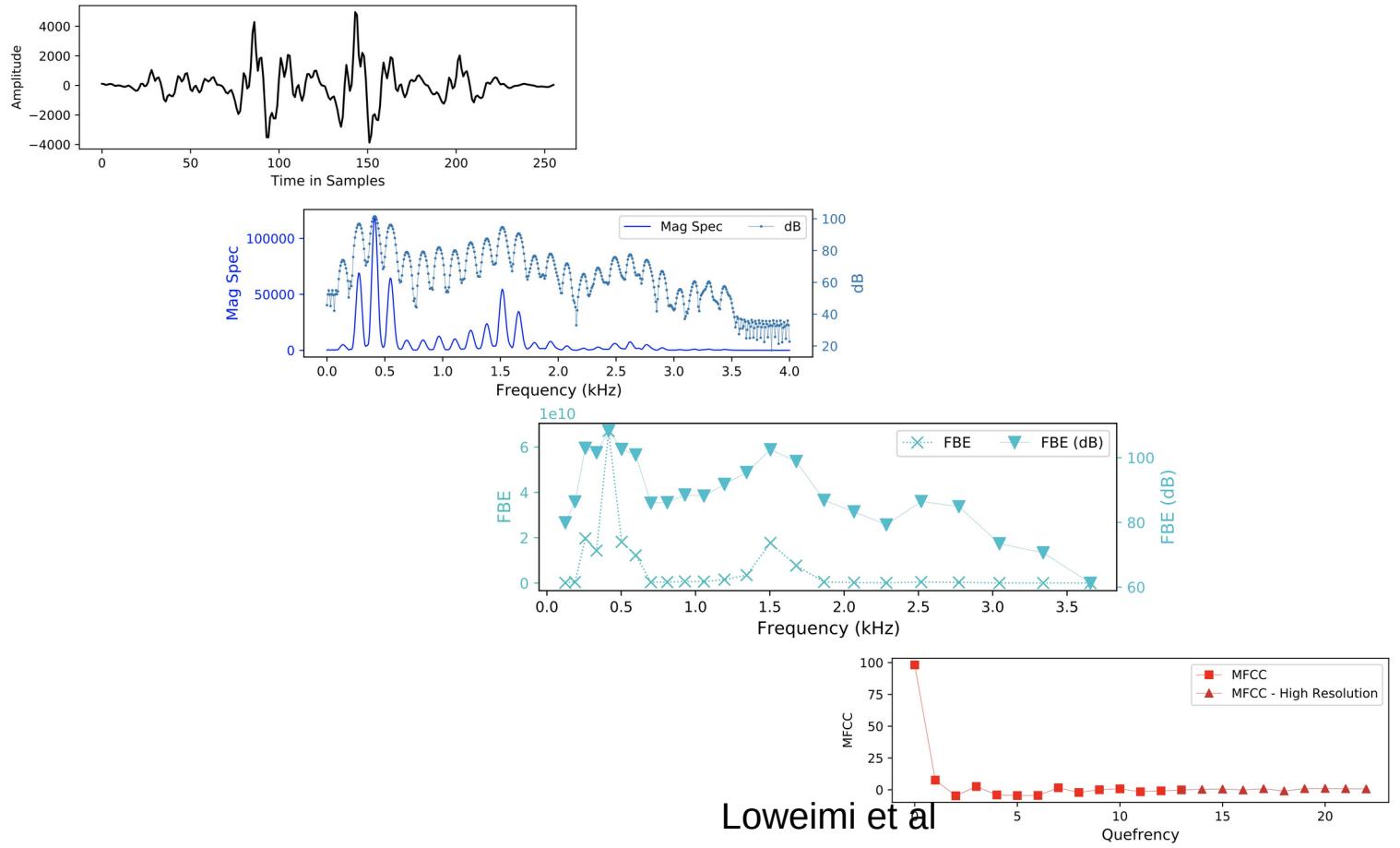
# Building A Perfect System Requires ...

- Perfect Filtering  $\leftrightarrow$  Architecture+Data+Training
- Perfect Match  $\leftrightarrow$  Architecture+Data+Training
- Perfect Input  $\leftarrow$  Raw signal representation ...

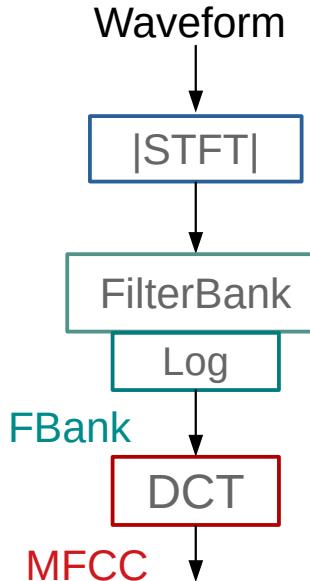
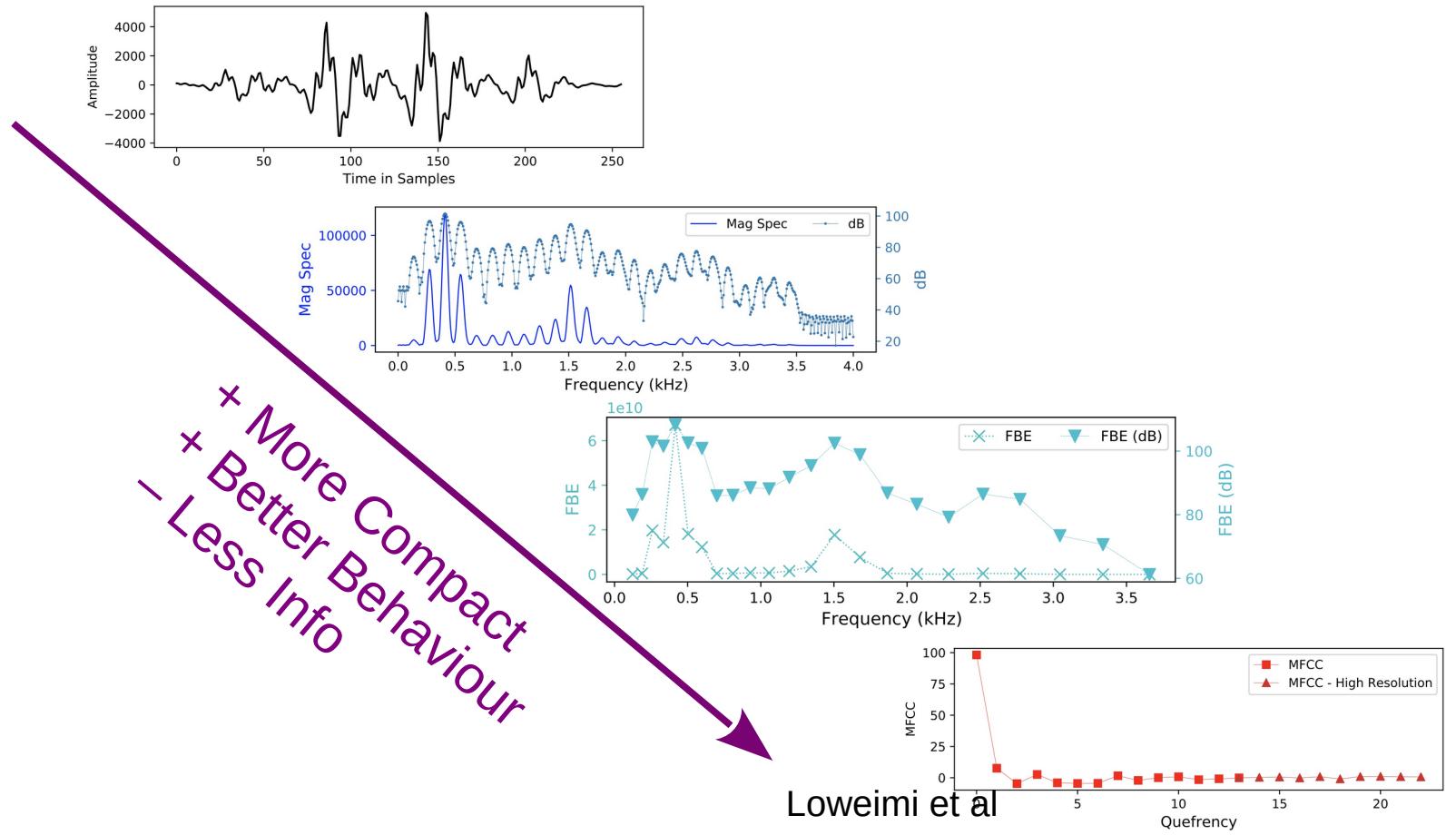
# Feature Extraction Pipeline ...



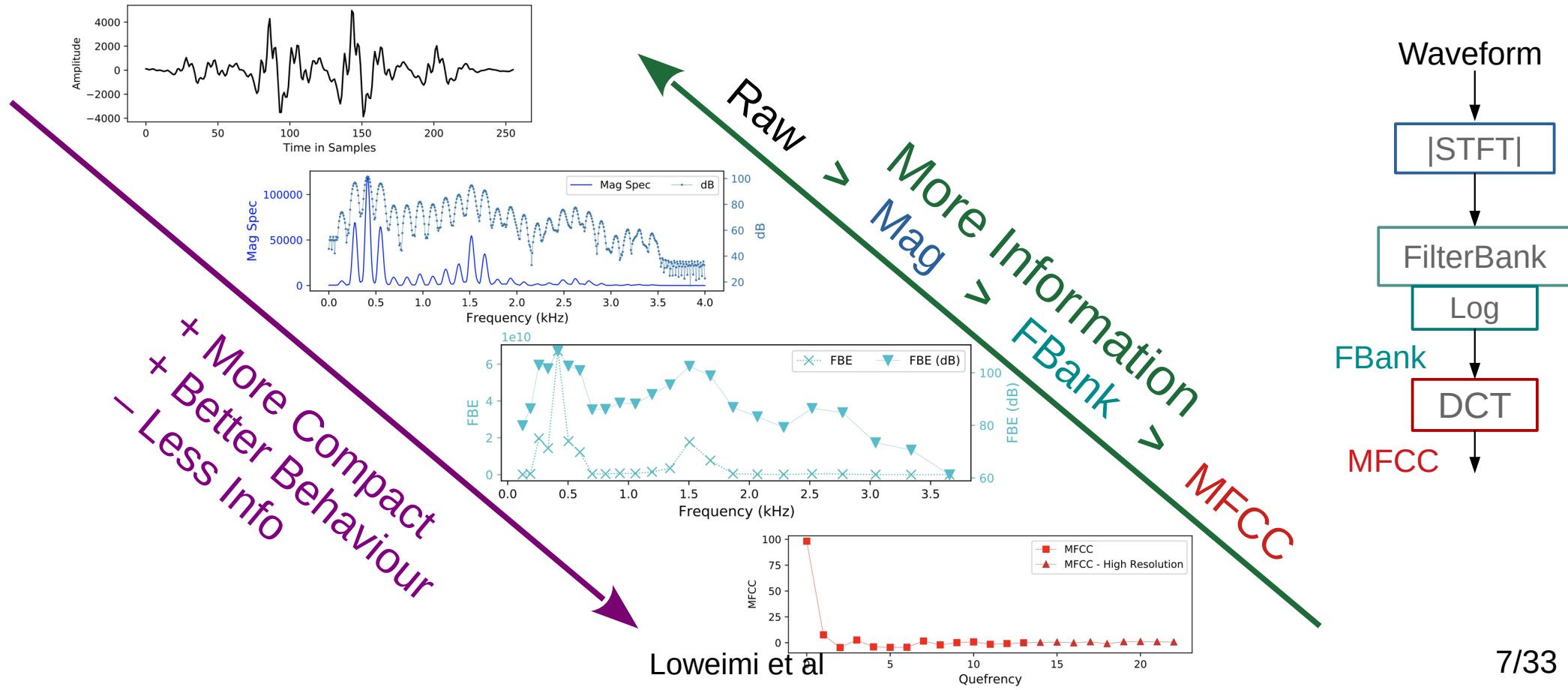
# Feature Extraction Pipeline ...



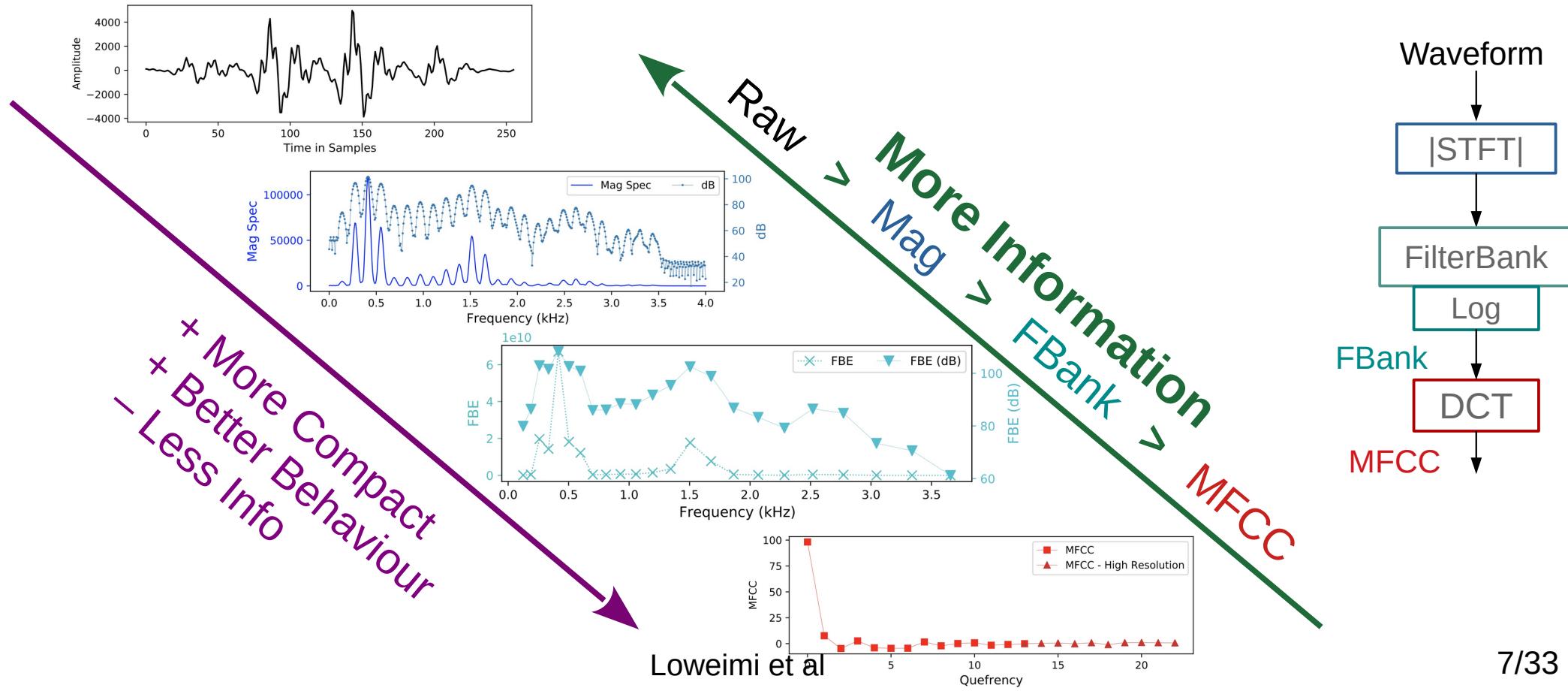
# Feature Extraction Pipeline ...



# Feature Extraction Pipeline ...

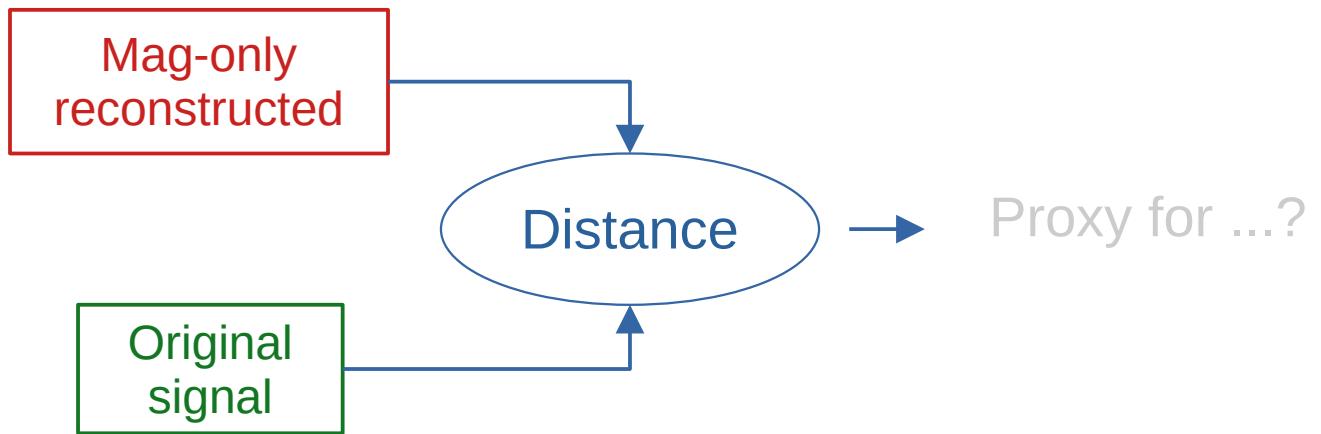


# Feature Extraction Pipeline ...

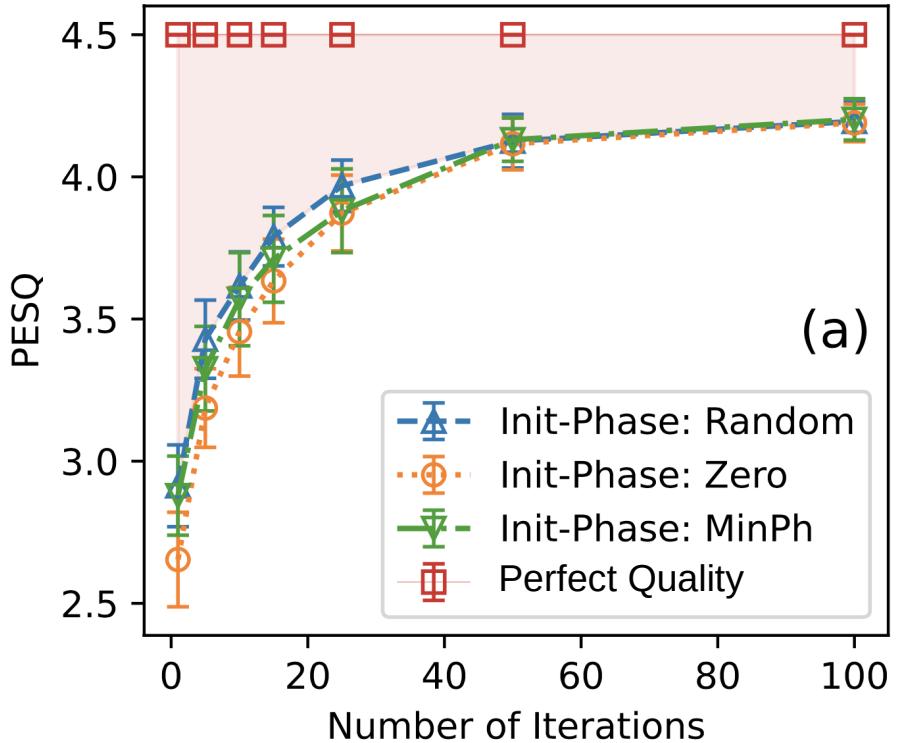


# Mag-Only Signal Reconstruction\*

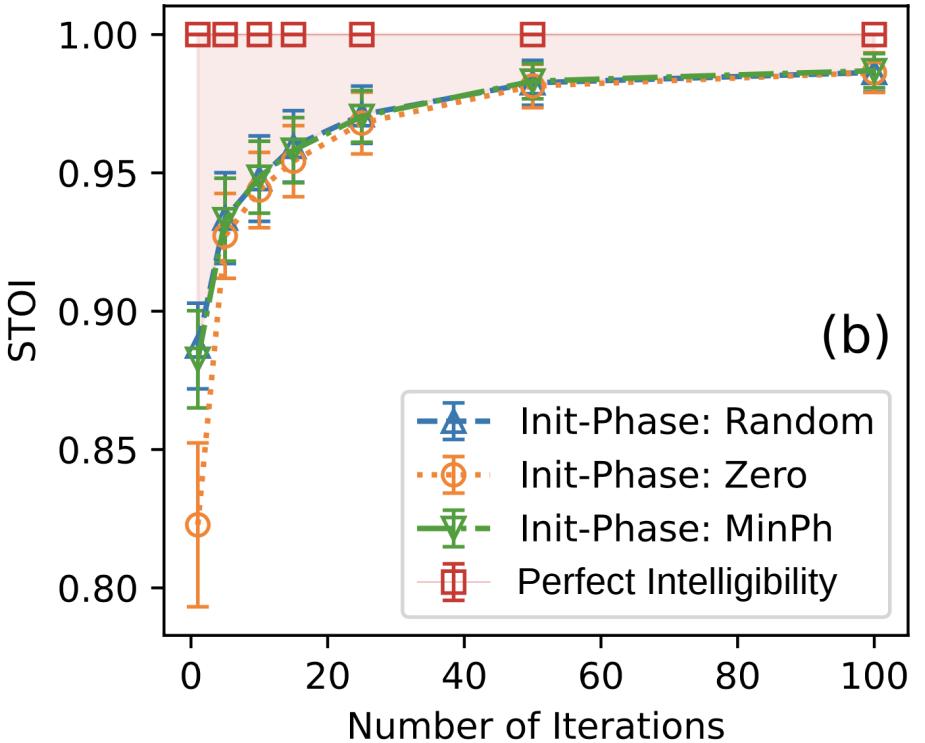
- ... proxy for magnitude info content



# Mag-Only Signal Reconstruction\*



(a)



(b)

PESQ  $\leftrightarrow$  QualitySTOI  $\leftrightarrow$  Intelligibility

# Outline

- Motivation
- Architecture
- Variants, Analysis & Interpretation
- Conclusion

# Signal Information Distribution

$$\begin{aligned}\mathbb{I}_{\text{signal}} &= \boxed{\mathbb{I}_{\text{waveform}}} = \mathbb{I}_{\text{Real}} \cup \mathbb{I}_{\text{Imag}} \\ &= \boxed{\mathbb{I}_{\text{Mag}} \cup \mathbb{I}_{\text{Phase}}} \\ &= \mathbb{I}_{\text{Mag}} \cup \mathbb{I}_{\text{All-Pass}} \\ &= \mathbb{I}_{\text{Mag}} \cup \mathbb{I}_{\text{Sign}} \\ &= \boxed{\mathbb{I}_{\text{Min-Ph}} \cup \mathbb{I}_{\text{All-Pass}}}\end{aligned}$$

Time Domain

Frequency Domain

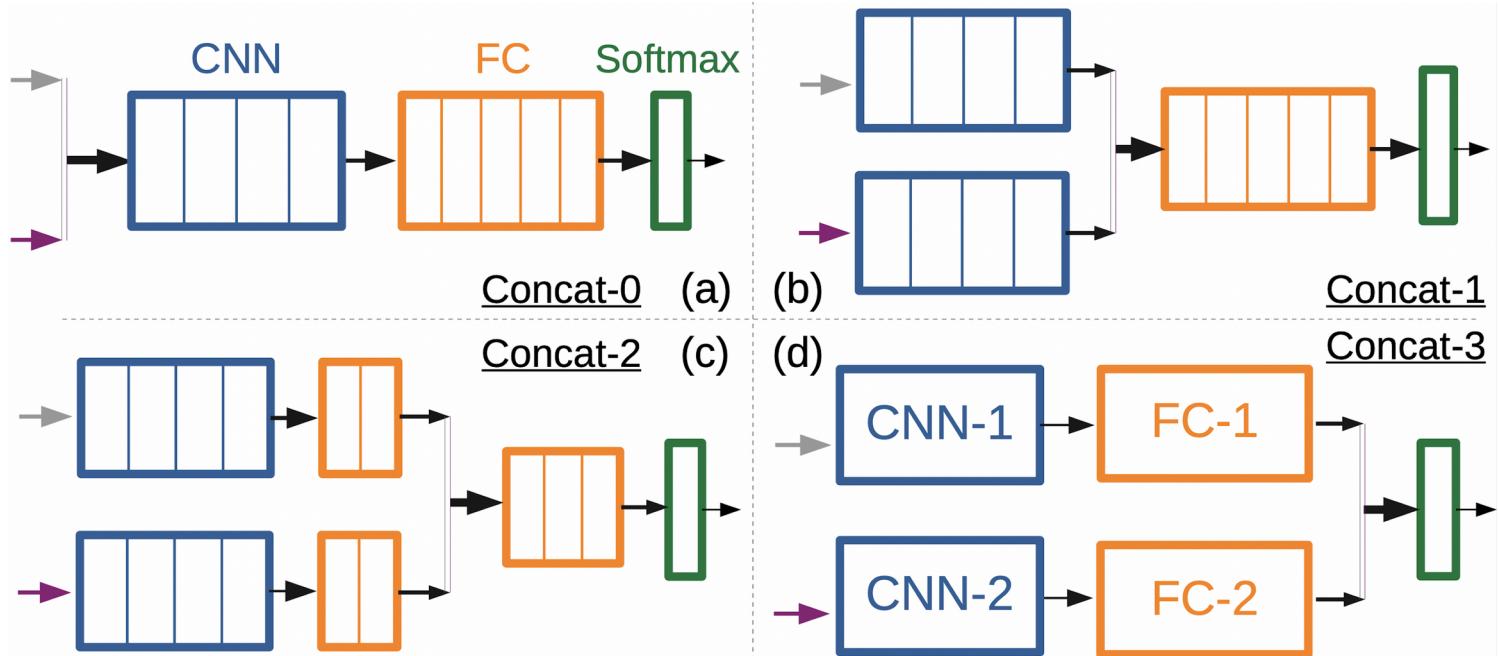
# Signal Information Distribution

$$\begin{aligned}\mathbb{I}_{\text{signal}} &= \mathbb{I}_{\text{waveform}} &= \mathbb{I}_{\text{Real}} \cup \mathbb{I}_{\text{Imag}} \\ &= \mathbb{I}_{\text{Mag}} \cup \mathbb{I}_{\text{Phase}} \\ &= \mathbb{I}_{\text{Mag}} \cup \mathbb{I}_{\text{All-Pass}} \\ &= \mathbb{I}_{\text{Mag}} \cup \mathbb{I}_{\text{Sign}} \\ &= \mathbb{I}_{\text{Min-Ph}} \cup \mathbb{I}_{\text{All-Pass}}\end{aligned}$$

Single-stream

Multi-stream

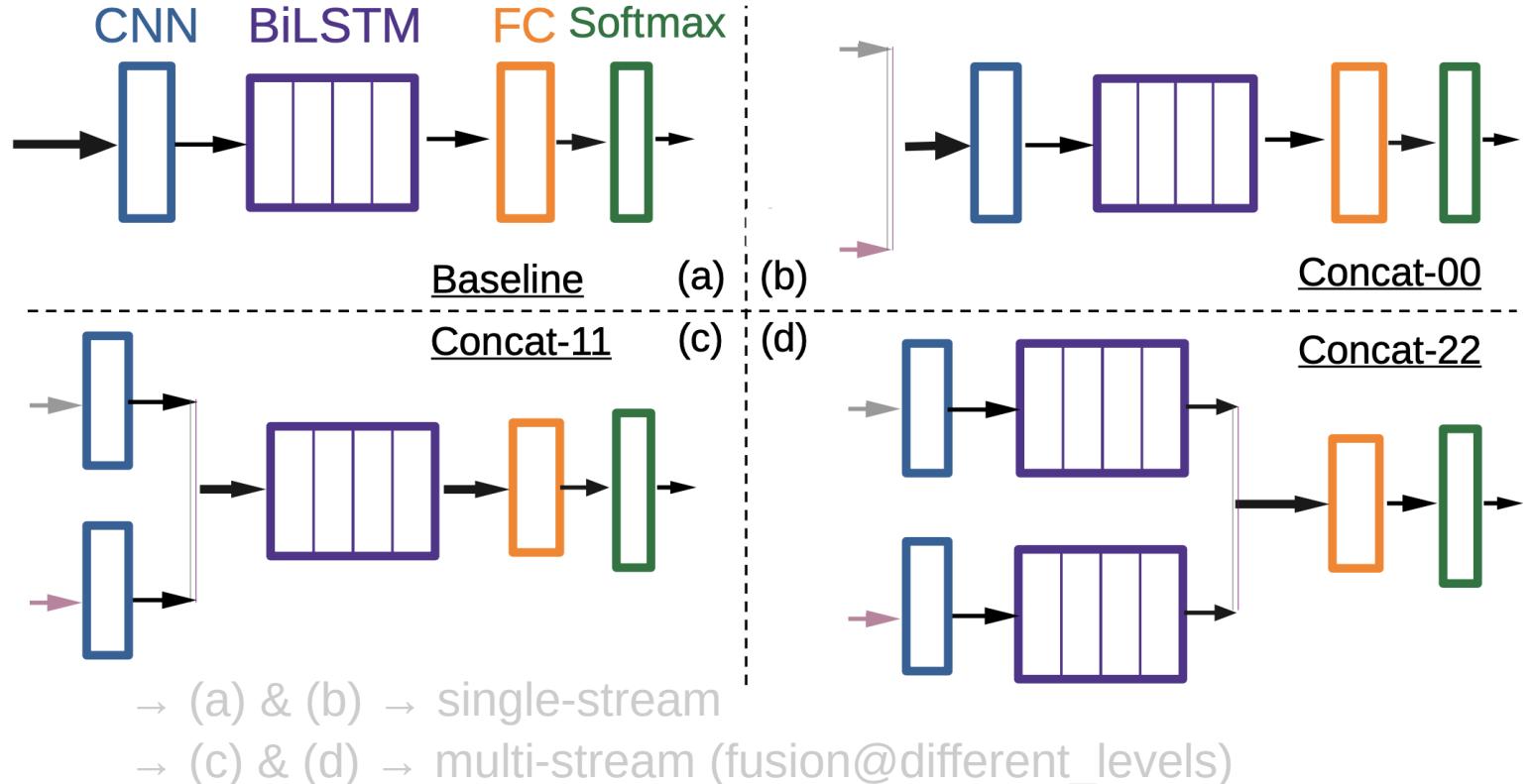
# Single- & Multi-stream Processing



→ (a) → single-stream (fusion @ input\_level)

→ (b), (c) & (d) → multi-stream (fusion@different\_levels)

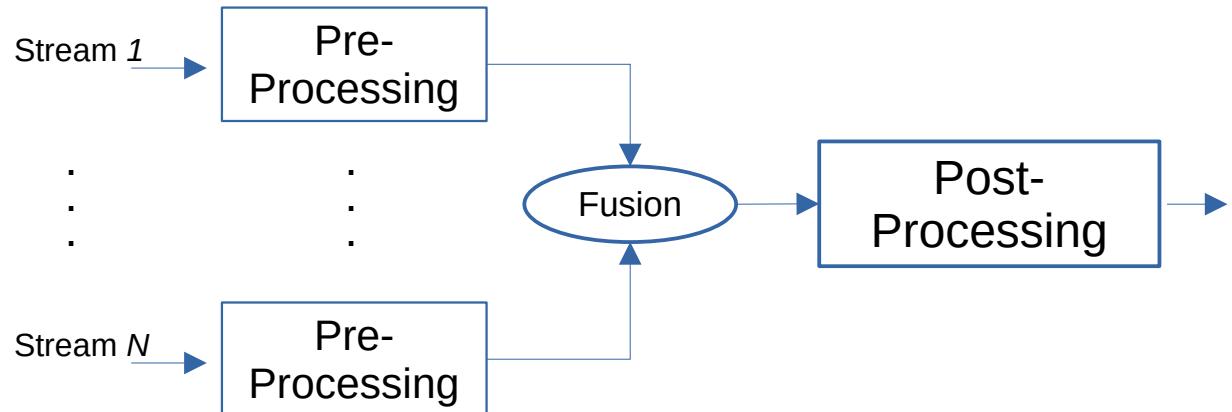
# Single- & Multi-stream Processing



# Multi-stream Proc. Advantages (1)

- Stream-specific pre-processing ... multi-modal inputs ...

e.g., CogMHear AVSEC



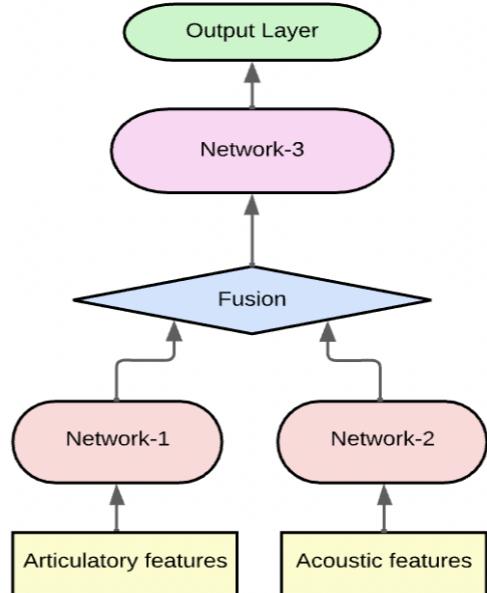
Zhengjun Yue<sup>1</sup>, Erfan Loweimi<sup>2</sup>, Zoran Cvetkovic<sup>2</sup>, Heidi Christensen<sup>1</sup> and Jon Barker<sup>1</sup><sup>1</sup> Department of Computer Science, University of Sheffield, UK<sup>2</sup> Department of Engineering, King's College London, UK

Under Review

# Acoustic-articulatory Multimodal Speech Recognition for Dysarthric Speech

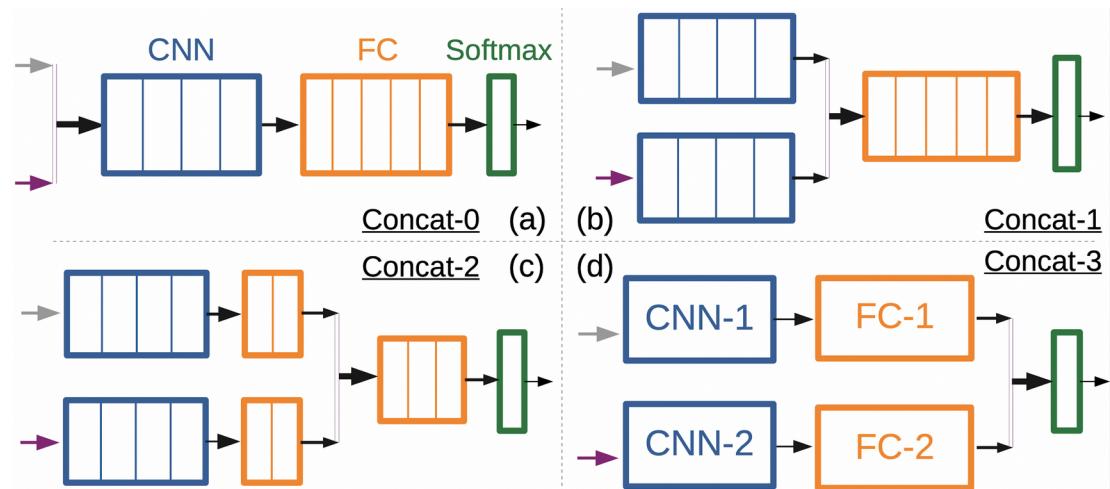
Zhengjun Yue (Member, IEEE), Erfan Loweimi (Member, IEEE), Zoran Cvetkovic (Senior Member, IEEE), Jon Barker (Member, IEEE), Heidi Christensen (Member, IEEE)

Input features	Systems	Severe		M/S		Moderate		Mild		Average	
		M01	M02	M04	M05	F03	F04	M03	Dys	Typ	
FBank83	baseline	60.4	55.2	68.7	46.6	31.5	16.8	8.4	35.6	11.7	
FBank83+Lip_EuD	concat-0	58.5	53.0	<b>66.7</b>	46.3	31.7	16.6	8.3	35.1	11.4	
FBank83+Lip_EuD	concat-1	<b>52.4</b>	56.2	69.4	<b>43.5</b>	<b>30.6</b>	15.9	7.5	34.4	10.8	
FBank83+Lip_EuD	concat-2	56.6	<b>52.9</b>	67.9	45.4	31.1	<b>14.5</b>	<b>7.6</b>	<b>34.3</b>	<b>10.4</b>	
FBank83+Lip_EuD	concat-3	93.5	99.3	98.6	93.1	64.5	31.0	24.5	64.4	33.1	
MFCC	baseline+	64.4	66.6	73.1	58.9	36.7	18.2	9.6	40.9	15.4	
MFCC+Lip_EuD	concat-2	56.6	55.2	70.5	47.1	34.8	16.1	9.8	37.9	12.6	



# Multi-stream Proc. Advantages (2)

- Fusion @ optimal level ... Trade-off ...
  - Higher levels → #param ↑, pre-proc. ↑, post-proc. ↓



# Outline

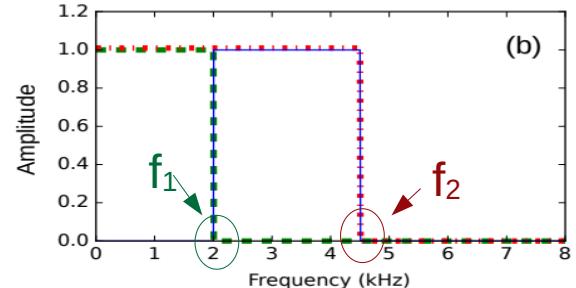
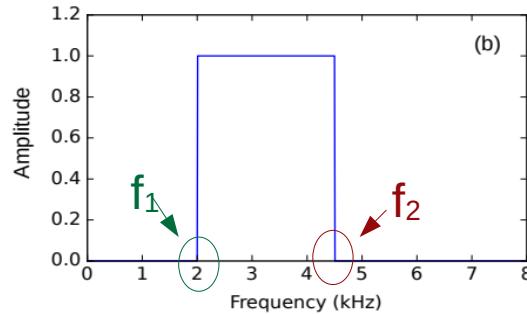
- Motivation
- Architectures
- Variants, Analysis & Interpretation
- Conclusion

# Raw Waveform Modelling by SincNet

- SincNet  $\leftrightarrow$  Parametric CNNs

$$h(t; \theta^{(i)}) = 2f_2^{(i)} \text{sinc}(2f_2^{(i)}t) - 2f_1^{(i)} \text{sinc}(2f_1^{(i)}t)$$

$$H(f; \theta^{(i)}) = \Pi\left(\frac{f}{2f_2^{(i)}}\right) - \Pi\left(\frac{f}{2f_1^{(i)}}\right)$$



# Raw Waveform Modelling by SincNet

- SincNet  $\leftrightarrow$  Parametric CNNs

$$h(t; \theta^{(i)}) = 2f_2^{(i)} \text{sinc}(2f_2^{(i)}t) - 2f_1^{(i)} \text{sinc}(2f_1^{(i)}t)$$

$$h(t; \theta^{(i)}) = \frac{1}{\pi t} (\sin(2\pi f_2^{(i)}t) - \sin(2\pi f_1^{(i)}t))$$

$$\sin \alpha - \sin \beta = 2 \sin \frac{\alpha - \beta}{2} \cos \frac{\alpha + \beta}{2}$$

$$h^{(i)}(t) = 2B^{(i)} \text{sinc}(B^{(i)}t) \cos(2\pi f_c^{(i)}t)$$

$$B^{(i)} = f_2^{(i)} - f_1^{(i)} \quad , \quad f_c^{(i)} = \frac{f_1^{(i)} + f_2^{(i)}}{2}$$

# Raw Waveform Modelling by SincNet

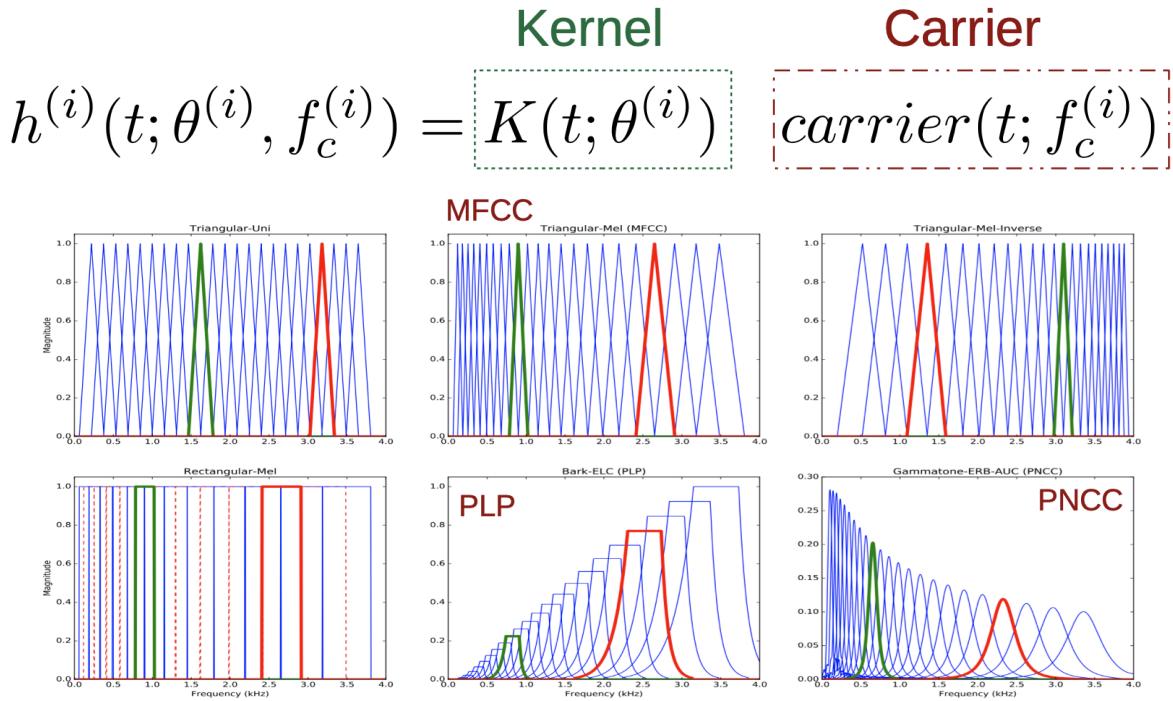
- SincNet  $\leftrightarrow$  Parametric CNNs

Kernel              Carrier

$$h^{(i)}(t) = [2B^{(i)} \operatorname{sinc}(B^{(i)}t)] \cos(2\pi f_c^{(i)}t)$$

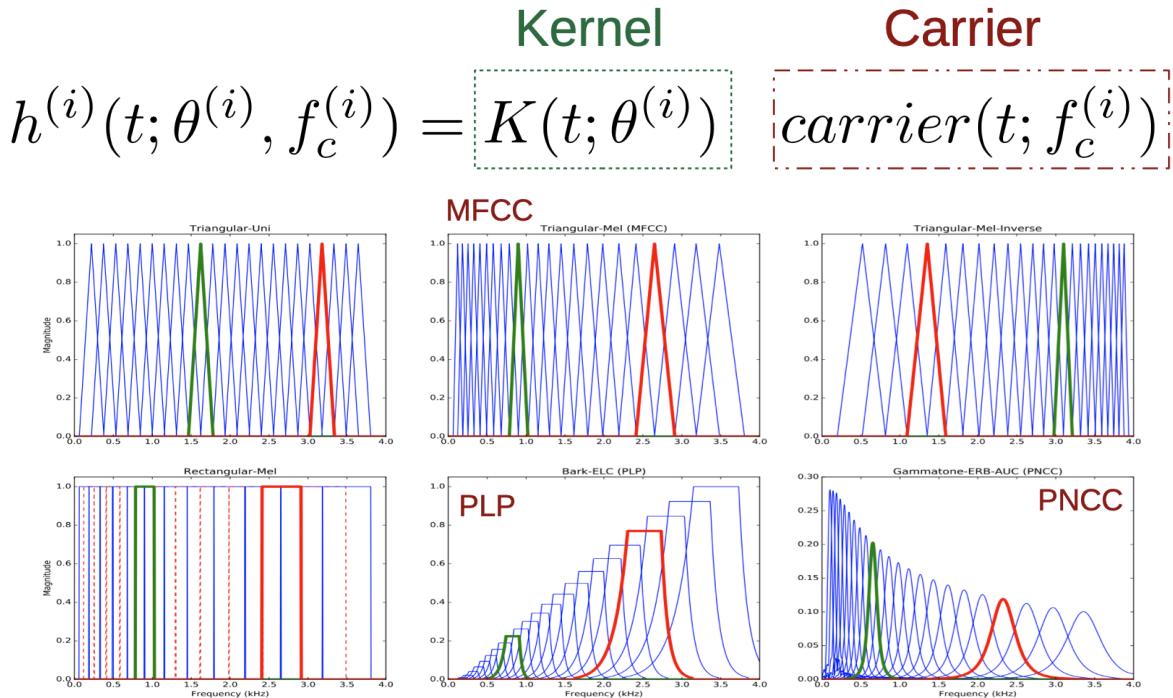
# Raw Waveform Modelling by XNet

- Parametric CNNs → Impose prior w/ perceptual flavour

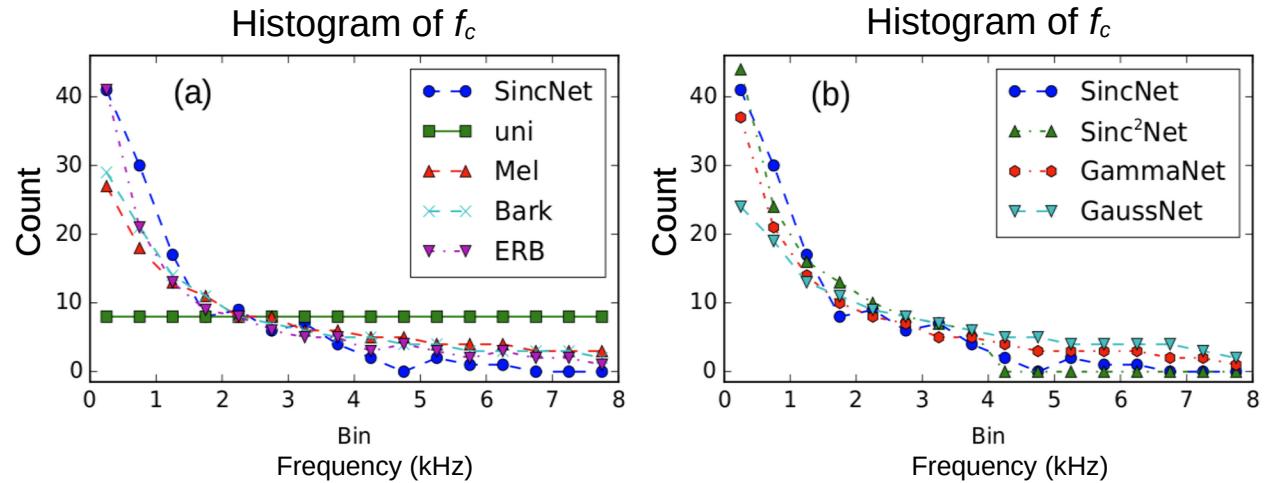


# Raw Waveform Modelling by XNet

- Parametric CNNs → Sinc<sup>2</sup>Net, GammNet, GaussNet

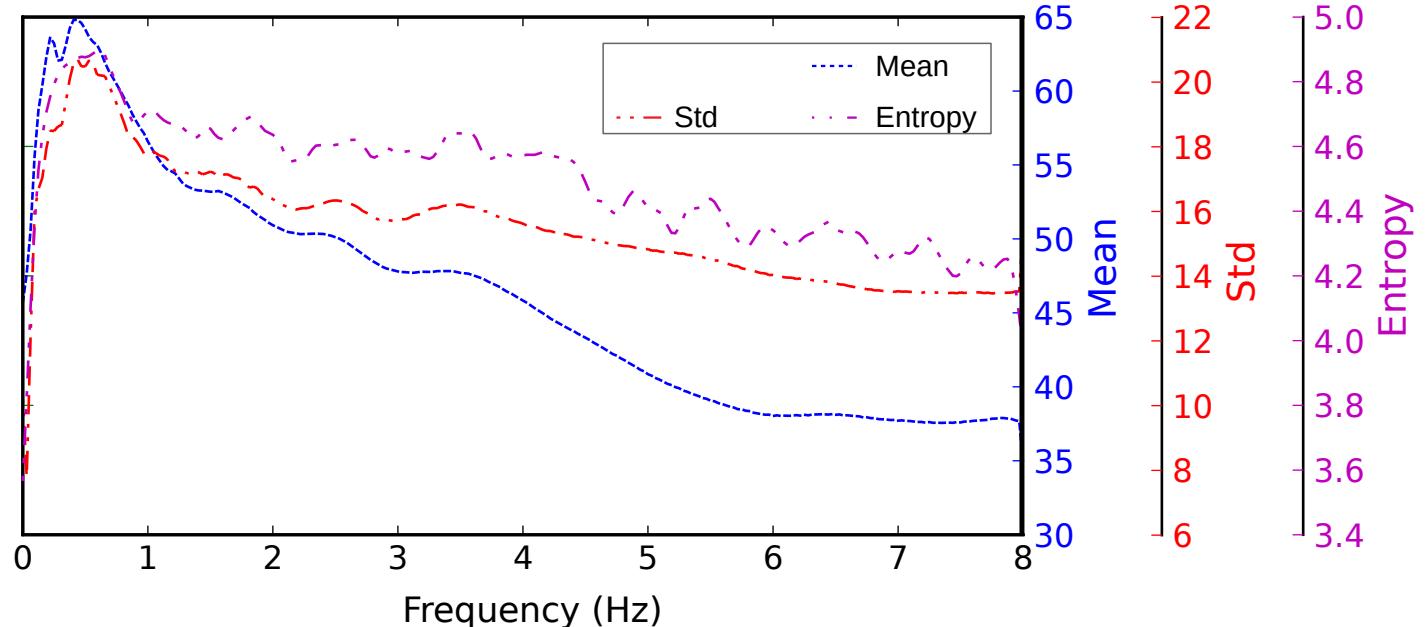


# Model Interpretation (1)

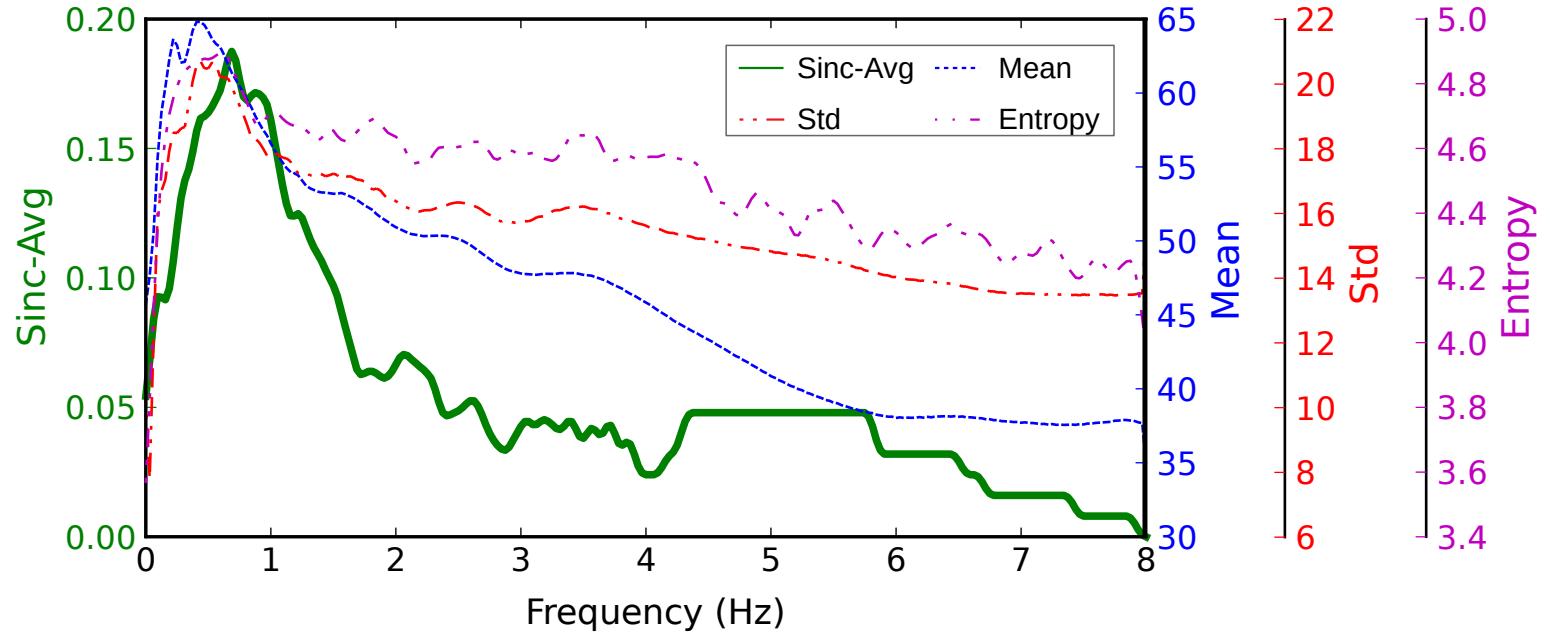


Filters are more discriminative & selective at lower frequencies.

# Model Interpretation (2)

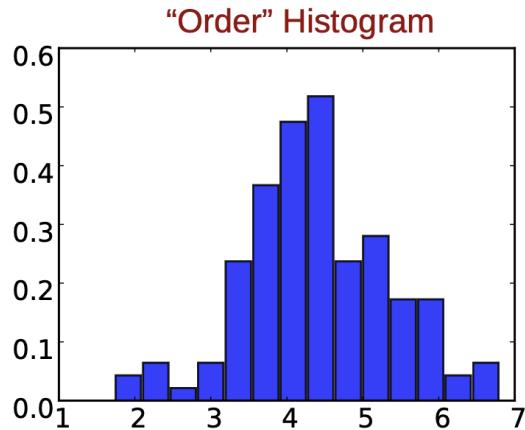


# Model Interpretation (2)



Model attends more important parts of data ...

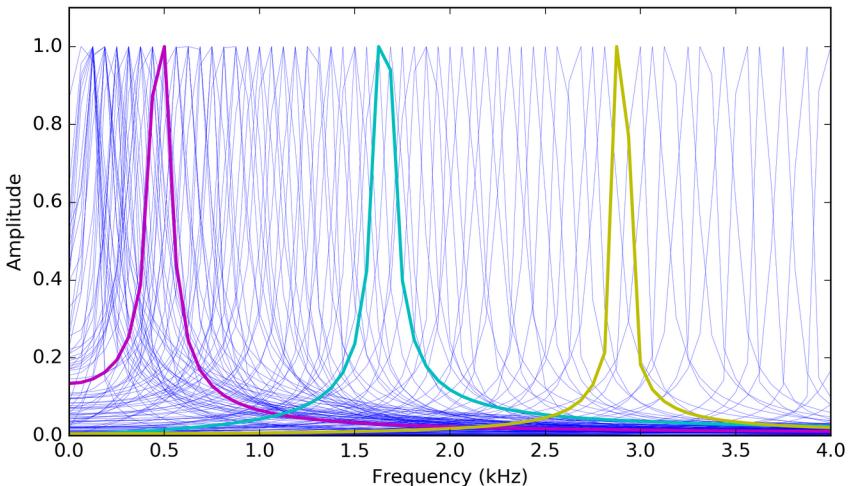
# Model Interpretation (3)



## A. A Comparison of Roex and Gammatone Amplitude Spectra

Schofield (1985) has recently demonstrated that a gammatone filter with order 4 provides a good fit to the average auditory filters presented in Patterson (1976).

$$K(t; \theta^{(i)}) = A^{(i)} t^{(N^{(i)} - 1)} e^{-2\pi B^{(i)} t}$$



# For more detail please refer to ...

## On Learning Interpretable CNNs with Parametric Modulated Kernel-based Filters

*Erfan Loweimi, Peter Bell and Steve Renals*

Centre for Speech Technology Research (CSTR), School of Informatics, University of Edinburgh

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



## Dysarthric Speech Recognition From Raw Waveform with Parametric CNNs

*Zhengjun Yue<sup>1,2,†</sup>, Erfan Loweimi<sup>1,3,†</sup>, Heidi Christensen<sup>2</sup>, Jon Barker<sup>2</sup>, Zoran Cvetkovic<sup>1</sup>*

<sup>1</sup> Department of Engineering, King's College London, UK

<sup>2</sup> Speech and Hearing Group (SPandH), University of Sheffield, UK

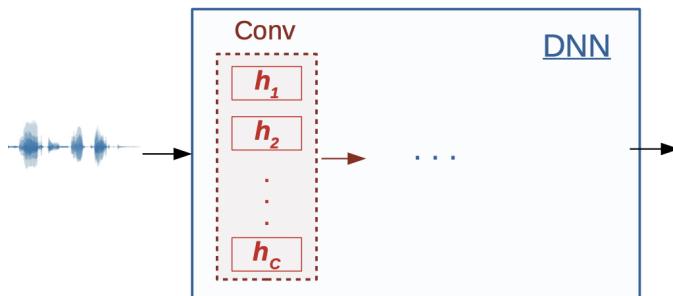
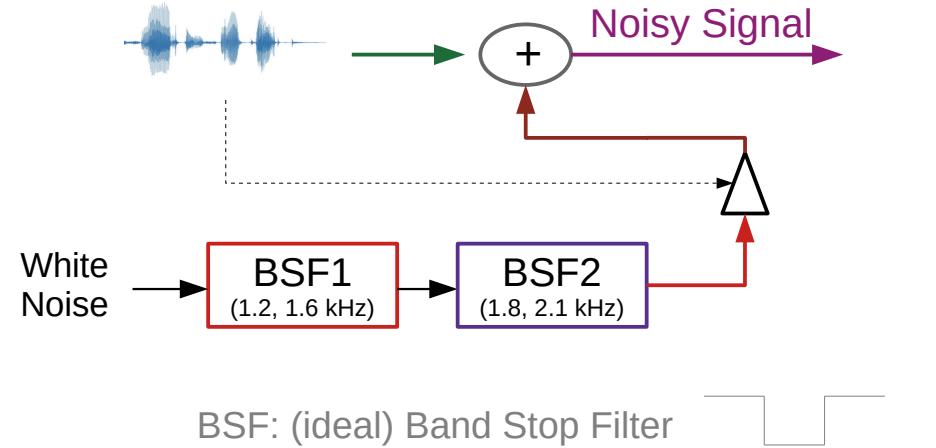
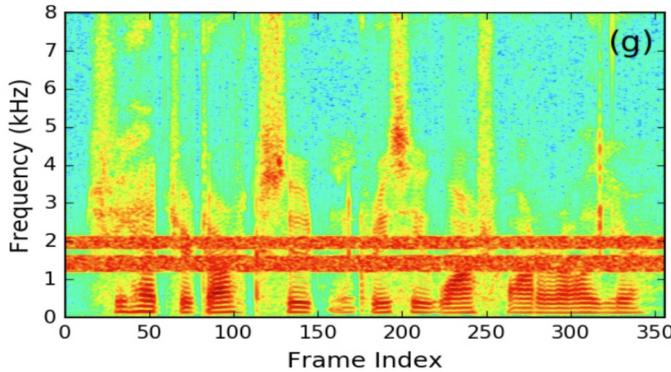
<sup>3</sup> Centre for Speech Technology Research (CSTR), University of Edinburgh, UK

{zhengjun.yue,erfan.loweimi,zoran.cvetkovic}@kcl.ac.uk,  
{heidi.christensen,j.p.barker}@sheffield.ac.uk}



# Raw Waveform Acoustic Modelling (2)

Task: TIMIT, Training data

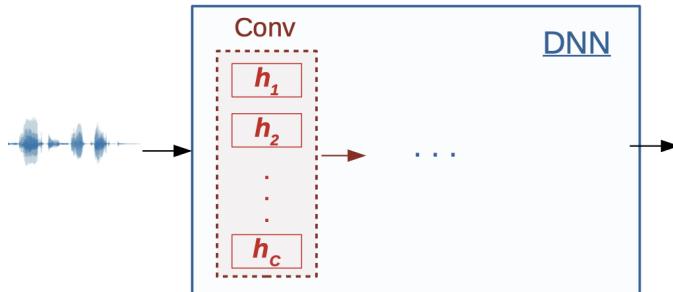
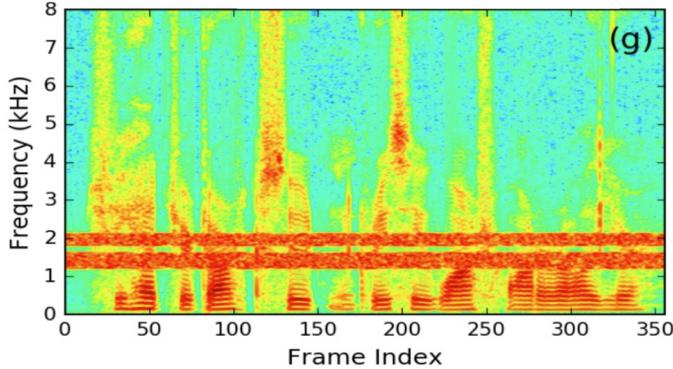


Average Frequency Response (AFR)

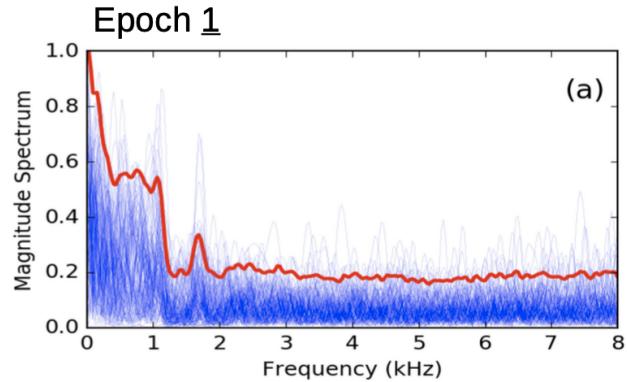
$$\text{AFR} = \frac{1}{C} \sum_{c=1}^C |H_c(\omega)|$$

# Raw Waveform Acoustic Modelling (2)

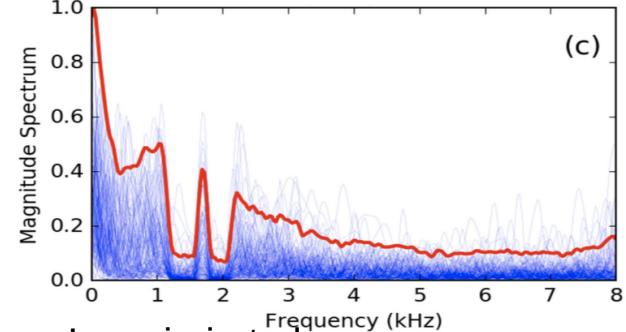
Task: TIMIT, Training data



Average Frequency Response (AFR)

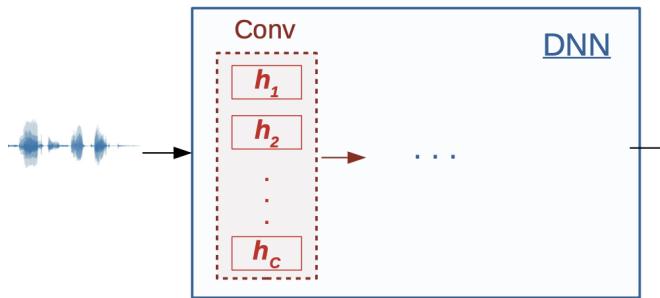
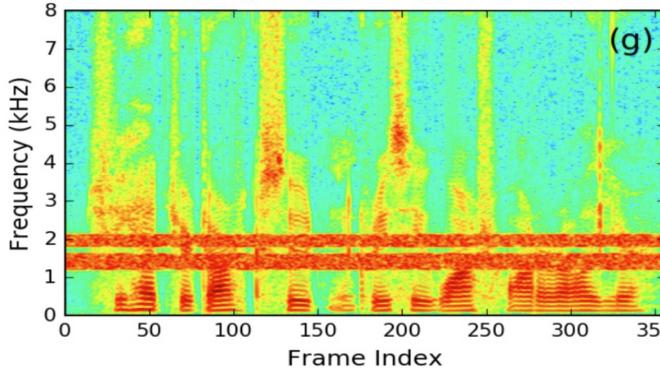


Epoch 20

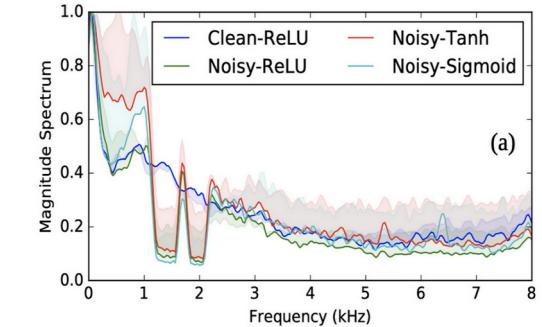
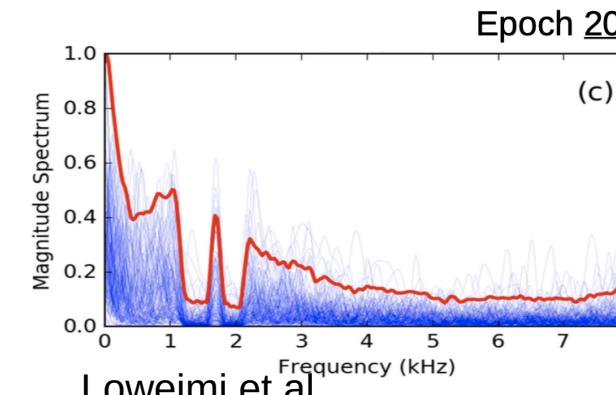
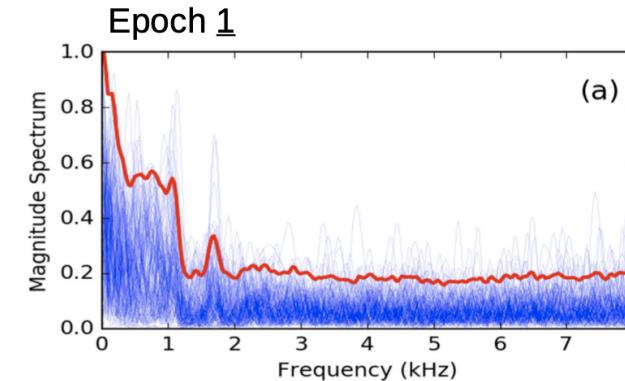


# Raw Waveform Acoustic Modelling (2)

Task: TIMIT, Training data



Average Frequency Response (AFR)



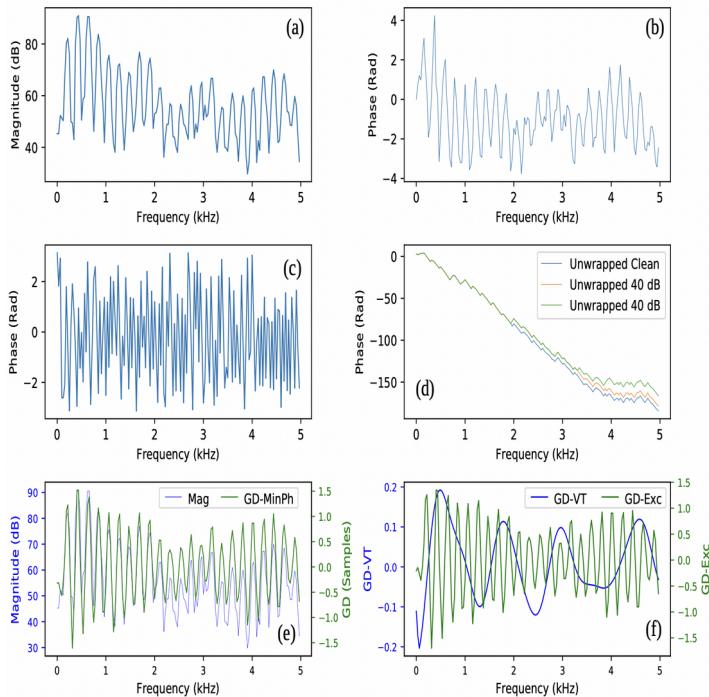
- 1) Enhancement  $\leftrightarrow$  ASR
- 2) Gradient vanishing ...

# SPEECH ACOUSTIC MODELLING FROM RAW PHASE SPECTRUM

Erfan Loweimi<sup>1</sup>, Zoran Cvetkovic<sup>2</sup>, Peter Bell<sup>1</sup> and Steve Renals<sup>1</sup>

<sup>1</sup> Centre for Speech Technology Research (CSTR), University of Edinburgh, UK

<sup>2</sup> Department of Engineering, King's College London, UK



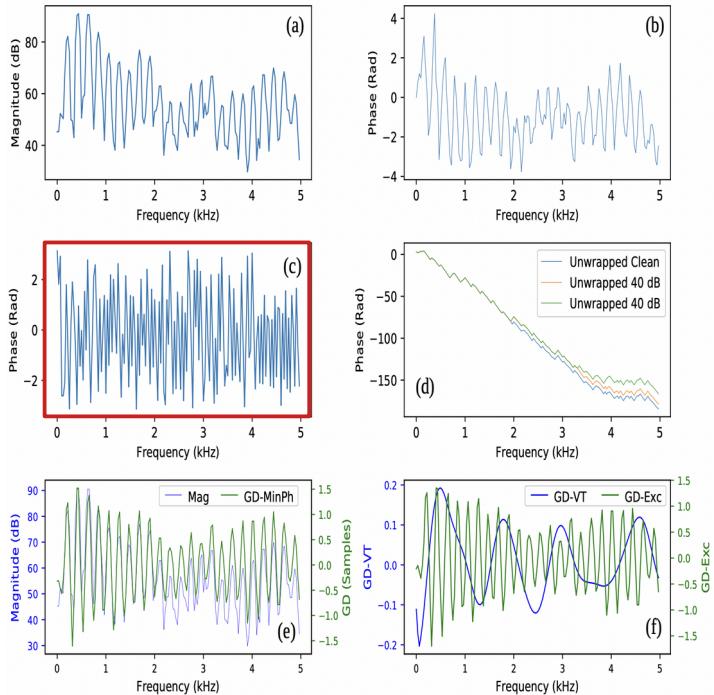
**Table 1.** TIMIT PER for different front-ends.

	Dev	Eval
MFCC	17.1	18.6
FBank	16.3	18.2
Mag	16.8	17.8
Mag <sup>0.1</sup>	15.9	17.6
Phase-Wrapped	21.6	23.7
Phase-UnWrapped	29.6	31.8
Phase-MinPh	16.8	18.6
GD-MinPh	16.9	18.4
GD-VT	18.2	19.3
GD-Exc	31.3	32.3
Concat-0	16.8	18.4
Concat-1	16.3	18.1
Concat-2	16.2	18.0
Concat-3	17.0	18.4

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

	Dev-93	Eval-92	Eval-93
MFCC	10.4	6.8	10.4
FBank	9.1	5.9	8.8
Mag	9.3	5.9	9.1
Mag <sup>0.1</sup>	<b>8.8</b>	<b>5.5</b>	<b>9.0</b>
Phase-Wrapped	9.9	6.1	10.4
Phase-UnWrapped	13.1	8.9	16.4
Phase-MinPh	9.3	5.8	9.4
GD-MinPh	<b>8.3</b>	<b>5.1</b>	<b>7.8</b>
GD-VT	8.6	5.4	7.6
GD-Exc	12.2	8.5	13.2
Concat-0	8.2	4.9	7.8
Concat-1	<b>7.9</b>	<b>4.8</b>	<b>7.4</b>
Concat-2	8.1	4.8	7.7
Concat-3	8.2	5.0	8.1

## SPEECH ACOUSTIC MODELLING FROM RAW PHASE SPECTRUM

Erfan Loweimi<sup>1</sup>, Zoran Cvetkovic<sup>2</sup>, Peter Bell<sup>1</sup> and Steve Renals<sup>1</sup><sup>1</sup> Centre for Speech Technology Research (CSTR), University of Edinburgh, UK<sup>2</sup> Department of Engineering, King's College London, UK**Table 1.** TIMIT PER for different front-ends.

	Dev	Eval
MFCC	17.1	18.6
FBank	16.3	18.2
Mag	16.8	17.8
Mag <sup>0.1</sup>	15.9	17.6
Phase-Wrapped	21.6	23.7
Phase-UnWrapped	29.6	31.8
Phase-MinPh	16.8	18.6
GD-MinPh	16.9	18.4
GD-VT	18.2	19.3
GD-Exc	31.3	32.3
Concat-0	16.8	18.4
Concat-1	16.3	18.1
Concat-2	16.2	18.0
Concat-3	17.0	18.4

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

	Dev-93	Eval-92	Eval-93
MFCC	10.4	6.8	10.4
FBank	9.1	5.9	8.8
Mag	9.3	5.9	9.1
Mag <sup>0.1</sup>	<b>8.8</b>	<b>5.5</b>	<b>9.0</b>
Phase-Wrapped	9.9	6.1	10.4
Phase-UnWrapped	13.1	8.9	16.4
Phase-MinPh	9.3	5.8	9.4
GD-MinPh	<b>8.3</b>	<b>5.1</b>	<b>7.8</b>
GD-VT	8.6	5.4	7.6
GD-Exc	12.2	8.5	13.2
Concat-0	8.2	4.9	7.8
Concat-1	<b>7.9</b>	<b>4.8</b>	<b>7.4</b>
Concat-2	8.1	4.8	7.7
Concat-3	8.2	5.0	8.1

# Dysarthric Speech Recognition, Detection and Classification using Raw Phase and Magnitude Spectra

Zhengjun Yue<sup>†</sup>, Erfan Loweimi<sup>†</sup> and Zoran Cvetkovic

Department of Engineering, King's College London, UK

{zhengjun.yue, erfan.loweimi, zoran.cvetkovic}@kcl.ac.uk



Table 1: WER of single-stream ADSR systems averaged over various severity levels (Mild, Mod: moderate, Sev: severe) on TORG0.

Model	3-L CNN						1-L CNN						
	Feature (Single-stream)	#Params (Millions)	Severity degrees			Average		#Params (Millions)	Severity degrees			Average	
			Sev	Mod	Mild	Dys	Typ		Sev	Mod	Mild	Dys	Typ
FBank		10.1	57.4	44.0	15.8	43.3±5.1	14.3±1.9	11.6	48.4	30.8	10.3	34.4±2.0	10.7±0.7
Mag		9.8	48.1	32.4	11.2	35.7±3.4	11.1±1.2	15.6	42.6	27.1	9.6	30.4±2.8	9.7±0.3
Wrapped-Phase		9.8	106.0	98.4	98.4	102.5±6.5	98.3±3.9	15.6	75.3	64.9	32.0	61.4±5.6	37.8±5.2
Unwrapped-Phase		9.8	88.4	86.7	64.3	81.5±3.9	68.7±3.6	15.6	81.0	73.5	42.1	68.8±1.9	46.0±4.2
MinPhase		9.8	53.1	37.1	12.4	38.7±3.6	12.2±1.6	15.6	43.8	28.1	9.1	31.2±3.0	9.1±0.5

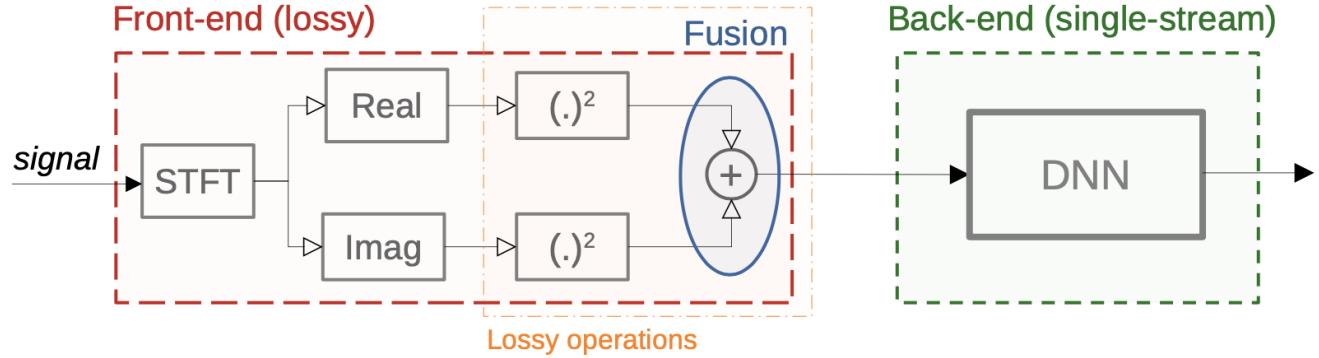
Table 2: WER of multi-stream ADSR systems averaged over various severity levels (Mild, Mod: moderate, Sev: severe) on TORG0.

Model	3-L CNN						1-L CNN						
	Feature (Multi-stream)	#Params (Millions)	Severity degrees			Average		#Params (Millions)	Severity degrees			Average	
			Sev	Mod	Mild	Dys	Typ		Sev	Mod	Mild	Dys	Typ
Mag+Mag		10.1	47.0	31.1	9.5	33.5±3.1	9.7±0.7	21.6	44.1	27.9	9.3	31.3±1.9	9.1±0.2
Mag+Mag+Mag		10.3	46.9	30.8	9.4	33.4±1.6	9.6±0.7	27.7	44.0	28.7	9.0	31.1±2.9	9.2±0.5
MinPhase+MinPhase		10.1	49.2	32.8	10.9	35.4±3.5	10.3±1.6	21.6	45.3	29.2	9.4	32.2±2.8	9.4±0.5
MinPhase+MinPhase+MinPhase		10.3	48.2	30.0	10.5	34.2±2.6	9.9±0.9	27.7	45.6	30.7	9.7	32.7±3.1	9.9±0.3
FBank+Cos(Phase)		10.2	51.2	33.8	12.1	36.9±3.2	11.5±1.1	17.7	50.8	35.9	11.5	37.0±1.4	12.1±0.9
FBank+MinPhase		10.2	47.3	30.9	11.5	34.2±2.6	10.6±1.0	17.7	44.7	29.1	10.1	32.1±2.6	10.4±0.6
FBank+Mag		10.2	46.8	30.5	10.8	33.6±2.1	10.6±0.9	17.7	43.7	28.1	10.2	31.4±2.4	10.0±0.7
Mag+WrappedPhase		10.1	47.9	31.5	10.1	34.2±3.0	10.0±1.9	21.6	48.4	33.7	10.9	35.2±2.1	10.9±1.5
Mag+Cos(Phase)		10.1	47.3	29.6	9.8	33.7±1.6	9.8±0.3	21.6	48.3	34.1	10.2	35.0±1.7	10.8±0.5
Mag+Sin(Phase)		10.1	48.9	30.8	10.2	34.7±3.4	10.1±0.8	21.6	48.6	32.4	10.2	34.8±2.5	10.5±0.4
Mag+MinPhase		10.1	48.4	31.7	10.5	34.6±2.6	10.4±1.5	21.6	44.2	28.4	9.2	31.4±2.5	9.0±0.2

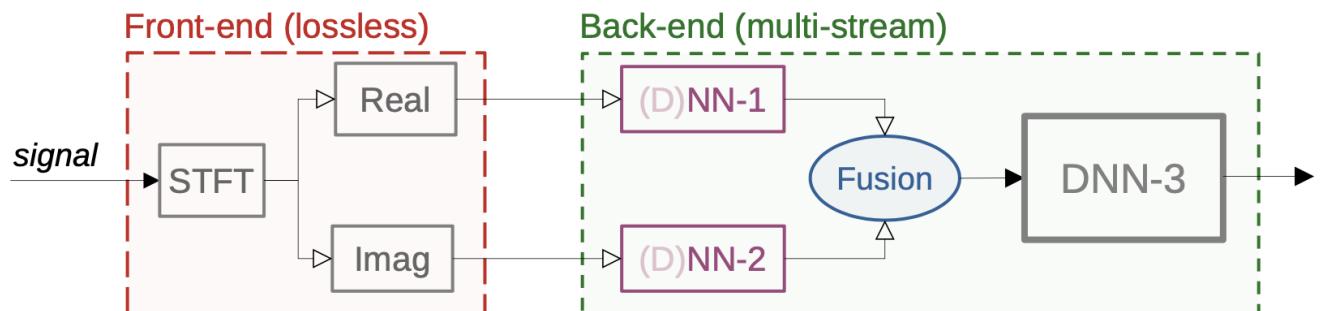
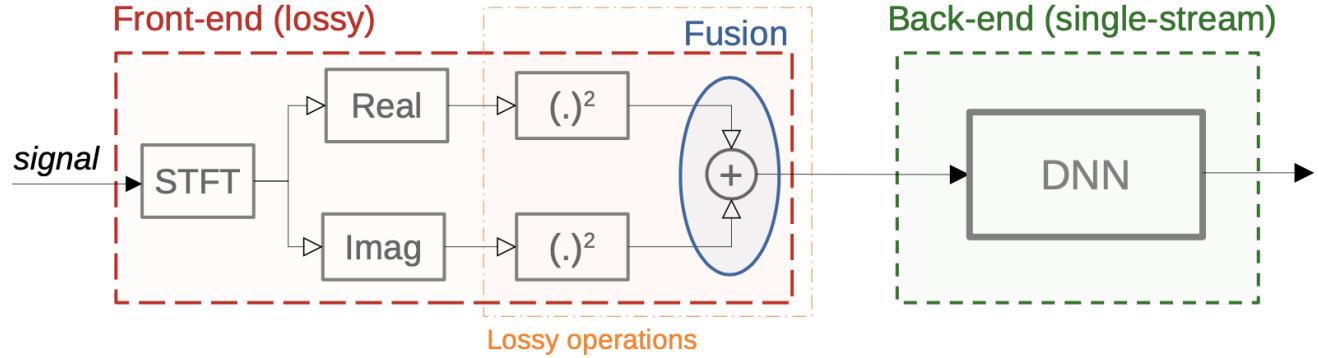
Table 4: WER of various ADSR systems on UASpeech.

Feature	FBank	Mag	MinPhase	Mag+MinPhase	[34]
UASpeech	31.7	30.4	30.8	<b>30.2</b>	30.5

# Mixing Learning & Signal Processing



# Mixing Learning & Signal Processing

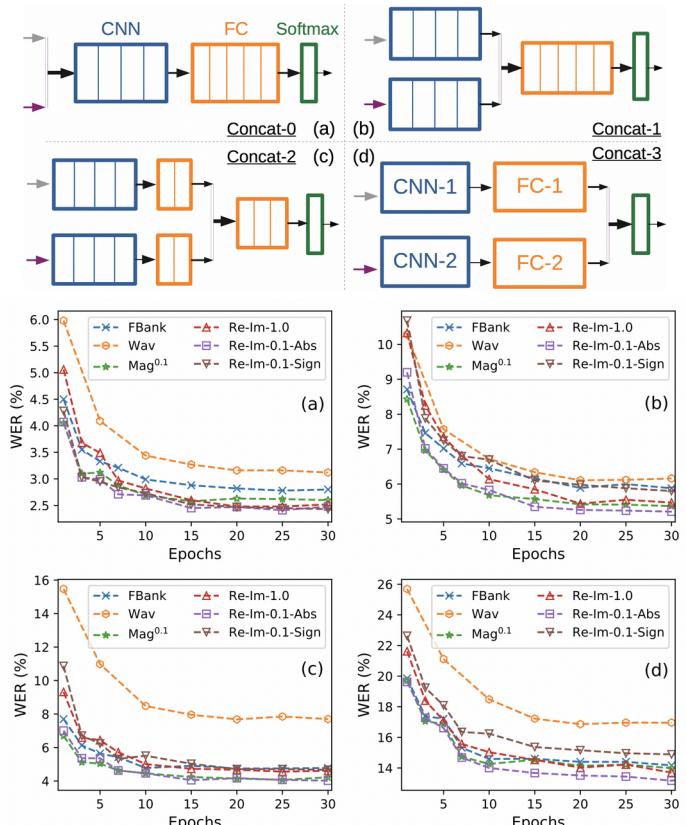


# Multi-stream Acoustic Modelling using Raw Real and Imaginary Parts of the Fourier Transform

Erfan Loweimi  (Member, IEEE), Zhengjun Yue  (Member, IEEE), Peter Bell  (Member, IEEE),  
Steve Renals  (Fellow, IEEE), Zoran Cvetkovic  (Senior Member, IEEE)

Feature	A	B	C	D	Avg
MFCC-clean-align	3.4	5.8	4.5	7.9	6.4
FBank-clean-align	2.8	5.1	3.2	6.3	5.3
Mag <sup>0.1</sup> -clean-align	2.7	4.7	3.3	5.8	4.9
Raw-wave-clean-align	2.7	<b>4.4</b>	4.0	6.4	5.1
Concat-0-0.1-Abs-clean-align	2.4	4.6	2.8	5.9	4.8
Concat-1-0.1-Abs-clean-align	2.4	4.5	2.9	5.7	4.7
Concat-2-0.1-Abs-clean-align	<b>2.3</b>	4.5	<b>2.5</b>	<b>5.6</b>	<b>4.6</b>
Concat-3-0.1-Abs-clean-align	2.5	4.8	3.0	6.2	5.1

- A: Clean
- B: Additive noise
- C: Channel noise
- D: Additive & channel noise



# Multi-stream Acoustic Modelling using Raw Real and Imaginary Parts of the Fourier Transform

Erfan Loweimi  (Member, IEEE), Zhengjun Yue  (Member, IEEE), Peter Bell  (Member, IEEE),  
Steve Renals  (Fellow, IEEE), Zoran Cvetkovic  (Senior Member, IEEE)

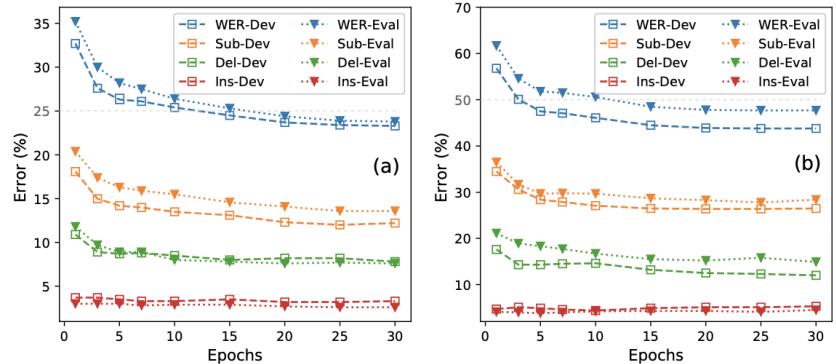
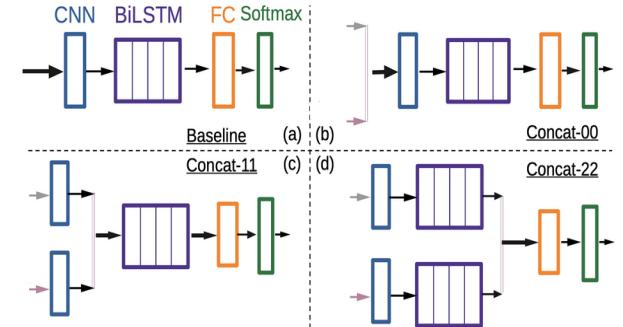
TABLE X

*WER on AMI-IHM and AMI-SDM (CLDNN).*

*Number of \* denotes number of convolutional layers. BS: batch size.*

	IHM		SDM	
	Dev	Eval	Dev	Eval
Raw-wave* (BS:8)	24.1	24.5	47.3	50.8
FBank* (BS:8)	23.8	24.4	44.2	48.1
Mag <sup>0.1*</sup> (BS:8)	23.4	24.3	43.8	47.8
Concat-11* (BS:4)	24.4	25.7	45.9	50.7
Concat-11* (BS:8)	24.0	25.1	45.2	49.7
Concat-11*-0.1-Abs (BS:4)	24.1	24.8	45.2	49.1
Concat-00*-0.1-Abs (BS:8)	23.9	24.3	<b>43.5</b>	47.6
Concat-11*-0.1-Abs (BS:8)	<b>23.3</b>	<b>23.8</b>	43.8	47.7
Concat-11**-0.1-Abs (BS:8)	23.4	24.2	43.7	<b>47.6</b>
Concat-11***-0.1-Abs (BS:8)	23.7	24.4	44.3	48.6
SAHR-Transformer (E2E) [69]	24.2	24.6	-	-
SAHR-Conformer (E2E) [69]	24.1	24.2	-	-
Multi-stream (E2E) [70]	-	-	-	54.9
Multi-scale Octave CNN (Hybrid) [71]	32.2	37.2	48.2	53.3
Parznet 2D-CNN (Hybrid) [72]	24.9	26.0	-	-
Parznet 2D-CNN+VI (Hybrid) [12]	24.7	25.7	-	-

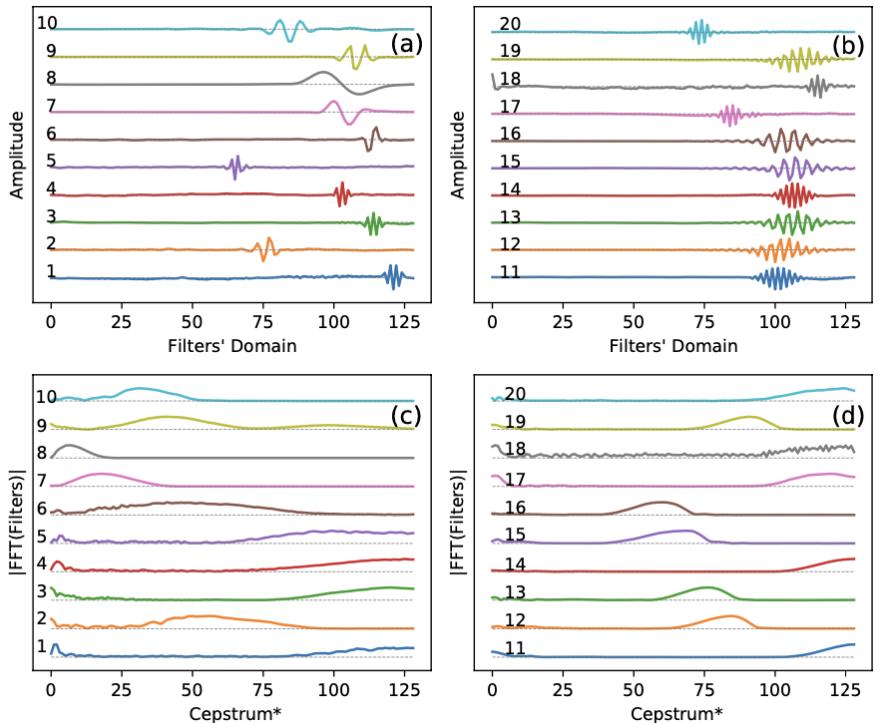
\* Studied tasks: TIMIT/NTIMIT, Aurora-4, WSJ, AMI, TORG



# Multi-stream Acoustic Modelling using Raw Real and Imaginary Parts of the Fourier Transform

Erfan Loweimi  (Member, IEEE), Zhengjun Yue  (Member, IEEE), Peter Bell  (Member, IEEE),  
 Steve Renals  (Fellow, IEEE), Zoran Cvetkovic  (Senior Member, IEEE)

Although CNNs are non-parametric,  
 some filters resemble parametric ones ...



# Speech Acoustic Modelling using Raw Source and Filter Components

Erfan Loweimi<sup>1</sup>, Zoran Cvetkovic<sup>2</sup>, Peter Bell<sup>1</sup>, and Steve Renals<sup>1</sup>

<sup>1</sup> Centre for Speech Technology Research (CSTR), University of Edinburgh, UK

<sup>2</sup> Department of Engineering, King's College London, UK

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

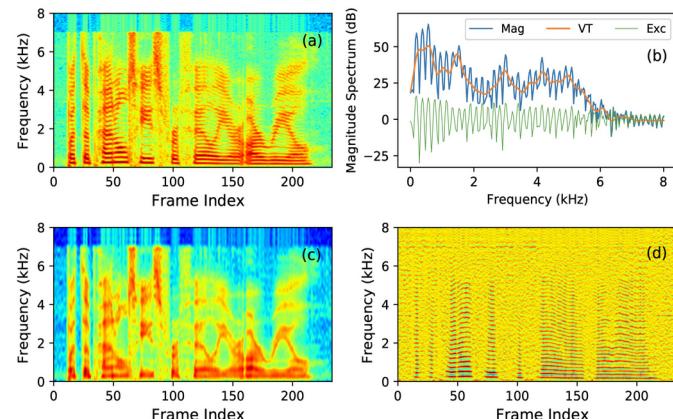
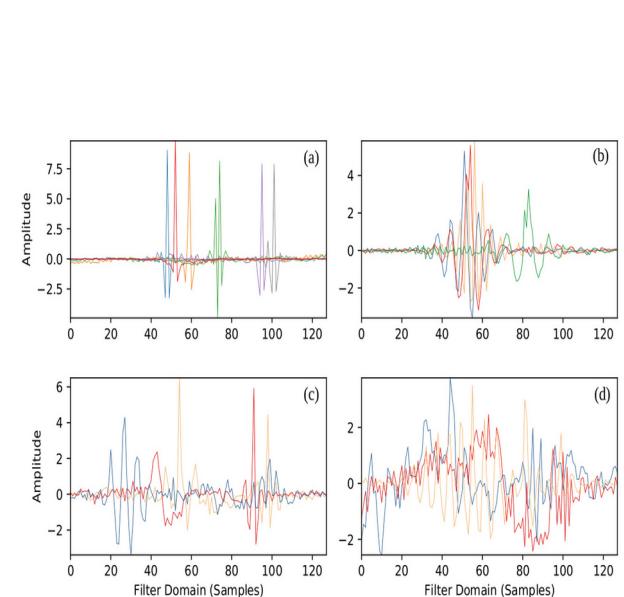


Table 2: WSJ WER for different front-ends.

	Dev	Eval-92	Eval-93
MFCC	10.4	6.8	10.4
FBank	9.1	5.9	8.8
Raw-wave	8.7	5.2	8.5
Mag <sup>0.1</sup> (baseline)	8.8	5.5	9.0
Exc	15.1	11.8	16.5
VT	9.6	6.3	9.1
Concat-1	7.9	4.5	7.5
Concat-2	7.9	4.6	7.6
Concat-3	8.1	4.8	7.6
Sinc-Concat-1	8.0	4.5	7.4

Table 3: Aurora-4 (multi-style) WER for different front-ends.

Feature	A	B	C	D	Avg
MFCC	3.5	6.8	7.1	16.5	10.7
FBank	2.9	5.9	4.5	14.5	9.2
Raw-wave	3.1	5.7	7.5	16.5	10.3
Mag <sup>0.1</sup> (baseline)	2.6	5.3	4.3	14.1	8.8
VT	3.0	6.0	5.1	15.0	9.6
Exc	6.4	15.8	16.2	32.6	22.4
Concat-1	2.4	5.1	4.1	13.0	8.2
Concat-2	2.5	5.2	4.3	13.3	8.4
Concat-3	2.5	5.5	4.5	13.9	8.8
Sinc-Concat-1	2.3	5.0	4.0	12.7	8.1



Zhengjun Yue<sup>1,2,†</sup>, Erfan Loweimi<sup>1,3,†</sup> and Zoran Cvetkovic<sup>1</sup>

- <sup>1</sup> Department of Engineering, King's College London, UK  
<sup>2</sup> Speech and Hearing Group (SPAndH), University of Sheffield, UK  
<sup>3</sup> Centre for Speech Technology Research (CSTR), University of Edinburgh, UK

Accepted

## Acoustic Modelling from Raw Source and Filter Components for Dysarthric Speech Recognition

Zhengjun Yue<sup>†</sup> (Member, IEEE), Erfan Loweimi<sup>†</sup> (Member, IEEE), Heidi Christensen (Member, IEEE),  
 Jon Barker (Member, IEEE), Zoran Cvetkovic (Senior Member, IEEE)

ITASLP

[UASpeech](#)

Training data	Feature	Dysarthric
Dys	FBank	43.1
Dys	VT+Exc	42.0
Dys	[17]	48.5
Both	FBank	42.9
Both	VT+Exc	42.2
Both-sp	FBank	31.7
Both-sp	VT+Exc	30.3
Both-sp	[24]	32.4
Both-sp	[52]	30.5

[TORGO](#)

System Feature	Average	Average
	Dysarthric	Typical
FBank	36.5	11.3
VT	36.6	11.2
VT+Exc	35.3	11.0
FBank + i-vector	36.3	11.0
VT + i-vector	36.0	11.0
VT+Exc + i-vector	35.4	10.8

Erfan Loweimi, Peter Bell and Steve Renals



Centre for Speech Technology Research (CSTR), School of Informatics, University of Edinburgh

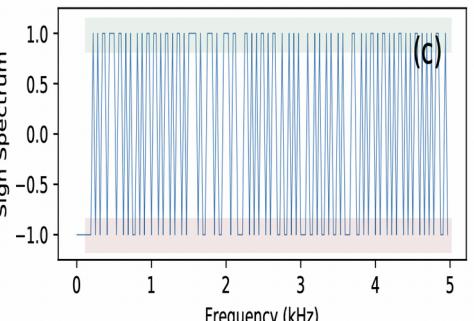
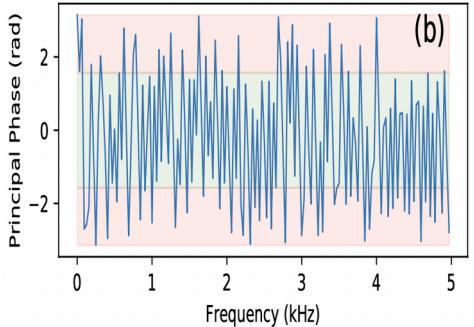
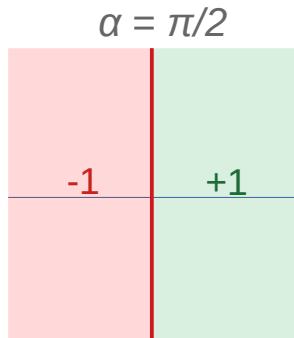
{e.loweimi, peter.bell, s.renals}@ed.ac.uk

IEEE TRANSACTIONS ON ACOUSTICS, SPEECH, AND SIGNAL PROCESSING, VOL. ASSP-31, NO. 5, OCTOBER 1983

# Signal Reconstruction from Signed Fourier Transform Magnitude

PATRICK L. VAN HOVE, MONSON H. HAYES, MEMBER, IEEE, JAE S. LIM, MEMBER, IEEE,  
AND ALAN V. OPPENHEIM, FELLOW, IEEE

$$S_X(\omega; \alpha) = \begin{cases} +1 & \alpha - \pi \leq \phi_X(\omega) \leq \alpha \\ -1 & \text{otherwise} \end{cases}$$





Centre for Speech Technology Research (CSTR), School of Informatics, University of Edinburgh

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

IEEE TRANSACTIONS ON ACOUSTICS, SPEECH, AND SIGNAL PROCESSING, VOL. ASSP-31, NO. 5, OCTOBER 1983

## Signal Reconstruction from Signed Fourier Transform Magnitude

PATRICK L. VAN HOVE, MONSON H. HAYES, MEMBER, IEEE, JAE S. LIM, MEMBER, IEEE,  
AND ALAN V. OPPENHEIM, FELLOW, IEEE

$$S_X(\omega; \alpha) = \begin{cases} +1 & \alpha - \pi \leq \phi_X(\omega) \leq \alpha \\ -1 & \text{otherwise} \end{cases}$$

$$\tilde{X}(\omega; \alpha) = S_X(\omega; \alpha) |X(\omega)|$$

	Hamming		Rectangular
	32 ms	512 ms	512 ms
Mag	4.22 ± 0.09	2.12 ± 0.24	2.38 ± 0.20
Mag+Sign	4.50 ± 0.00	4.20 ± 0.08	4.48 ± 0.02
Gain in PESQ	0.27	2.08	2.10



Centre for Speech Technology Research (CSTR), School of Informatics, University of Edinburgh

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

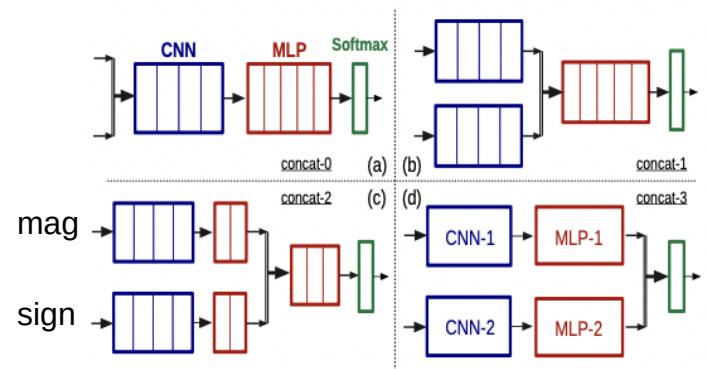
IEEE TRANSACTIONS ON ACOUSTICS, SPEECH, AND SIGNAL PROCESSING, VOL. ASSP-31, NO. 5, OCTOBER 1983

## Signal Reconstruction from Signed Fourier Transform Magnitude

PATRICK L. VAN HOVE, MONSON H. HAYES, MEMBER, IEEE, JAE S. LIM, MEMBER, IEEE,  
AND ALAN V. OPPENHEIM, FELLOW, IEEE

$$S_X(\omega; \alpha) = \begin{cases} +1 & \alpha - \pi \leq \phi_X(\omega) \leq \alpha \\ -1 & \text{otherwise} \end{cases}$$

$$\tilde{X}(\omega; \alpha) = \underline{S_X(\omega; \alpha)} |X(\omega)|$$



SUBMITTED TO IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING

# Phonetic Error Analysis Beyond Phone Error Rate

Erfan Loweimi  (Member, IEEE) , Andrea Carmantini  (Member, IEEE) , Peter Bell  (Member, IEEE) ,  
Steve Renals  (Fellow, IEEE) , Zoran Cvetkovic  (Senior Member, IEEE)

# Research Question

- Contribution of each **broad phonetic class** on PER?

PER@TestSet = 14.1%

$$PER = \sum_{\text{all classes}} PER_{class}$$

???

# Research Question

- Contribution of each **broad phonetic class** on PER?

PER@TestSet = 14.1%

$$PER = \sum_{\text{all classes}} PER_{\text{class}}$$

???

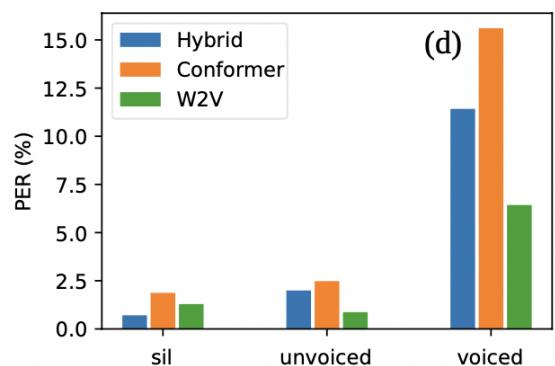
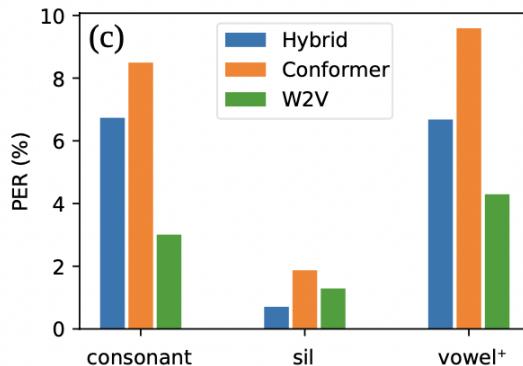
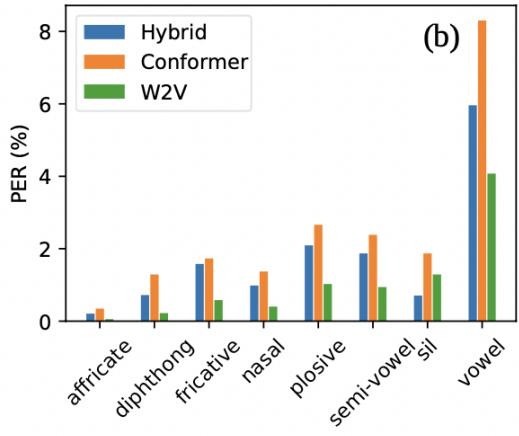
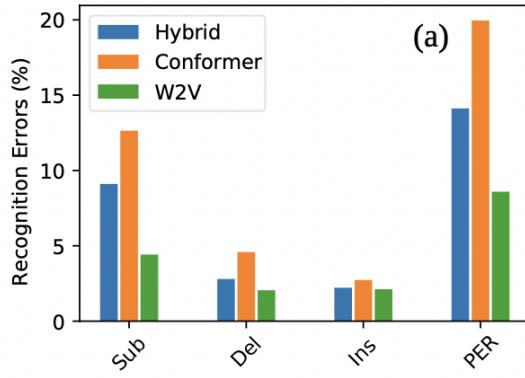
TABLE I  
*Mapping to the 8-class broad phonetic classes.*

classes	phones
Affricates	ch jh
Diphthongs	aw ay ey ow oy
Fricatives	ch f s sh th v z
Nasal	m n ng
Plosive	b d dx g k p t
Semi-vowel	hh l r w y
Vowel	aa ae ah eh er ih iy uh uw
Silence	sil

TABLE II  
*Mapping to the consonant, vowel<sup>+</sup>, silence, voiced and unvoiced BPCs.*

classes	phones
vowel <sup>+</sup>	aw ay ey ow oy aa ae ah eh er ih iy uh uw
Consonant	b ch d dh dx f g hh jh k l m n ng p r s sh t th v w y z
Silence	sil
Voiced	aa ae ah aw ay b d dh dx eh eer ey g hh ih iy jh l m n ng ow oy r uh uw v w y z
Unvoiced	ch f k p s sh t th

# Error Analysis



\* Most confused  
\* Second most confused

Loweimi et al

Wav2Vec 2.0 (pre-trained)  
[PER@test: 8.6]

Legend

- aff: affricate
- dip: diphthong
- fri: fricative
- nas: nasal
- plo: plosive
- sem: semi-vowel
- sil: silence
- vow: vowel
- con: consonant
- sil: silence
- vow<sup>+</sup>: vow+dip
- sil: silence
- unv: unvoiced
- voi: voiced

	aff	dip	fri	nas	plo	sem	sil	vow
True Label	0	0	0	0	2	1	0	0
	0	0	0	0	0	1	1	8
aff	0	0	0	0	2	1	0	0
dip	0	0	0	0	0	1	1	8
fri	1	0	37	0	2	1	2	0
nas	0	0	0	12	0	0	2	0
plo	2	0	2	0	19	0	1	0
sem	1	0	1	0	1	1	0	28
sil	0	0	1	0	8	1	0	0
vow	0	19	2	2	0	16	1	373

(a)

	aff	dip	fri	nas	plo	sem	sil	vow
True Label	0	0	0	0	2	1	0	0
	0	0	0	0	0	1	1	8
aff	0	0	0	0	2	1	0	0
dip	0	0	0	0	0	1	1	8
fri	1	0	37	0	2	1	2	0
nas	0	0	0	12	0	0	2	0
plo	2	0	2	0	19	0	1	0
sem	1	0	1	0	1	1	0	28
sil	0	0	1	0	8	1	0	0
vow	0	19	2	2	0	16	1	373

(b)

	aff	dip	fri	nas	plo	sem	sil	vow
True Label	0	0	0	0	2	1	0	0
	0	0	0	0	0	1	1	8
aff	0	0	0	0	2	1	0	0
dip	0	0	0	0	0	1	1	8
fri	1	0	37	0	2	1	2	0
nas	0	0	0	12	0	0	2	0
plo	2	0	2	0	19	0	1	0
sem	1	0	1	0	1	1	0	28
sil	0	0	1	0	8	1	0	0
vow	0	19	2	2	0	16	1	373

(c)

	aff	dip	fri	nas	plo	sem	sil	vow
True Label	0	0	0	0	2	1	0	0
	0	0	0	0	0	1	1	8
aff	0	0	0	0	2	1	0	0
dip	0	0	0	0	0	1	1	8
fri	1	0	37	0	2	1	2	0
nas	0	0	0	12	0	0	2	0
plo	2	0	2	0	19	0	1	0
sem	1	0	1	0	1	1	0	28
sil	0	0	1	0	8	1	0	0
vow	0	19	2	2	0	16	1	373

Confusion Matrix

# Towards Robust Waveform-Based Acoustic Models

Dino Oglic , Zoran Cvetkovic , Senior Member, IEEE, Peter Sollich , Steve Renals , Fellow, IEEE,  
and Bin Yu, Fellow, IEEE

**Goal:** time-domain data augmentation helps ...

TEST SAMPLE	CLEAN	CLEAN		CLEAN		CLEAN		CLEAN		CLEAN		CLEAN			
	BANDLIM.	-	NOTCH	-	WIDEPASS	WIDEPASS	RIR	-	GAUSS	GAUSS	RIR	-	NOTCH	WIDEPASS	
	ERR.	ERR.	DETR.	ERR.	DETR.	ERR.	DETR.	ERR.	DETR.	ERR.	DETR.	ERR.	DETR.	ERR.	DETR.
<b>A. SUMMARY OVER CLEAN SPEECH WITH TRAINING MICROPHONES</b>															
CLEAN	2.58	2.58	-	2.56	1%	2.58	-	2.47	4%	2.54	2%	<b>1.96</b>	24%		
<b>B. SUMMARY OVER NOISY SPEECH WITH TRAINING MICROPHONES</b>															
CAR	3.53	3.61	2%	4.04	14%	3.92	11%	3.89	10%	<b>3.36</b>	5%	15.92	351%		
BABBLE	6.58	9.86	50%	5.98	9%	<b>5.66</b>	14%	6.58	-	5.74	13%	16.66	153%		
RESTAURANT	7.64	8.16	7%	7.64	-	<b>7.38</b>	3%	8.41	10%	7.45	2%	15.99	109%		
STREET	<b>7.04</b>	7.23	3%	8.43	20%	7.83	11%	8.14	16%	7.32	4%	25.67	265%		
AIRPORT	<b>6.26</b>	9.21	47%	6.41	2%	6.44	3%	6.46	3%	6.63	6%	12.67	102%		
TRAIN	7.06	<b>6.65</b>	6%	8.16	16%	7.73	9%	8.89	26%	7.29	3%	26.56	276%		
B. AVERAGE	6.35	7.45	17%	6.78	7%	6.49	2%	7.06	11%	<b>6.30</b>	1%	18.91	198%		
<b>C. SUMMARY OVER CLEAN SPEECH WITH DIFFERENT MICROPHONES (UNSEEN DURING TRAINING)</b>															
CLEAN	7.79	7.58	3%	8.44	8%	12.29	58%	<b>7.49</b>	4%	8.03	3%	19.47	150%		
<b>D. SUMMARY OVER NOISY SPEECH WITH DIFFERENT MICROPHONES (UNSEEN DURING TRAINING)</b>															
CAR	10.46	<b>7.85</b>	25%	12.46	19%	15.32	46%	9.98	5%	8.78	16%	34.32	228%		
BABBLE	16.22	17.60	9%	16.59	2%	19.04	17%	18.70	15%	<b>15.37</b>	5%	38.93	140%		
RESTAURANT	18.16	<b>18.08</b>	-	19.63	8%	21.65	19%	20.79	14%	18.79	3%	36.52	101%		
STREET	18.74	<b>16.63</b>	11%	20.74	11%	22.73	21%	20.51	9%	17.21	8%	49.32	163%		
AIRPORT	<b>16.50</b>	17.37	5%	16.98	3%	20.31	23%	18.57	13%	16.66	1%	35.59	116%		
TRAIN	19.15	<b>16.44</b>	14%	20.94	9%	22.92	20%	21.32	11%	18.48	3%	48.07	151%		
D. AVERAGE	16.54	<b>15.66</b>	5%	17.89	8%	20.33	23%	18.31	11%	15.88	4%	40.46	145%		
<b>SUMMARY OVER ALL 14 TEST SAMPLES</b>															
AVERAGE	10.55	10.63	1%	11.36	8%	12.56	19%	11.59	10%	<b>10.26</b>	3%	26.98	156%		

# Outline

- Motivation
- Architectures
- Variants, Analysis & Interpretation
- Conclusion

# Conclusion

- Raw signal modelling avoid suboptimal info loss
- Amenable to model interpretation/analysis
- Imposing prior ... parametric CNN ... useful
- Still Spectral rep. are more robust than raw waveform
- Future work: data augmentation, self-supervised learning (leverage pre-trained models, e.g. wav2vec), ...

# That's it!

- Thank you for your attention!
- Q & A

SpeechWave

