Raw Sign and Magnitude Spectra for Multi-head Acoustic Modelling

Erfan Loweimi

Peter Bell and Steve Renals

Centre for Speech Technology Research (CSTR)

University of Edinburgh

Outline

- Motivation
- Signal Information Distribution
- Sign Spectrum
- Combination of Raw Magnitude & Sign Spectra
- Experimental Results
- Conclusion

Outline

- Motivation
- Signal Information Distribution
- Sign Spectrum
- Combination of Raw Magnitude & Sign Spectra
- Experimental Results
- Conclusion

Motivation

- Reviewers' Comments
- Components of A Perfect Information Processing System

Reviewers’ Comments ...

- *… I really enjoyed reading this paper … The approach is plausible and less ad hoc than much recent work ... dealing with phase …*
- *The paper shows that some good thinking and theory at the signal level can go hand in hand with a DNN ... w/o the need for blindly pumping tons of data ...*
- *… This paper provides a novel and strong contribution ...*
- *... is very well written, exhibit a clear structure and guides the reader nicely through the presentation of the topic …*
- *… is technically sound and the presented research well motivated …*
- *… it is an approach that is very worthwhile being shared at Interspeech ...*

Perfect Info Processing System (1)

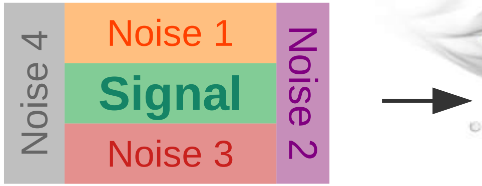
1) Perfect information filtering

- ONLY pass through the task-correlated info → *Discriminability*
- Filter the rest → *Robustness & Generalisation*

Perfect Info Processing System (1)

1) Perfect information filtering

- ONLY pass through the task-correlated info → *Discriminability*
- Filter the rest → *Robustness & Generalisation*



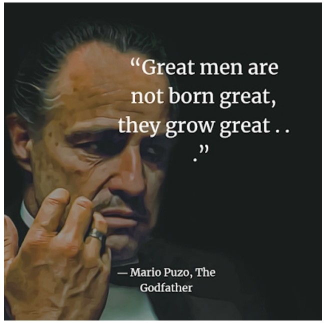
* Signal: task-correlated info
* Noise: task-irrelevant info

Perfect Info Processing System (1)

1) Perfect information filtering

- ONLY pass through the task-correlated info → *Discriminability*
- Filter the rest → *Robustness & Generalisation*

Task: ASR



→

→

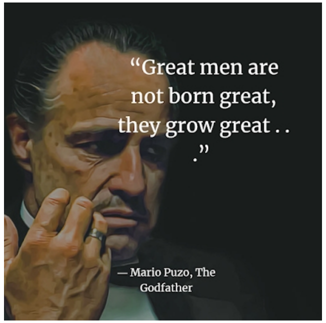
“Great men are not born great, they grow great”

Perfect Info Processing System (1)

1) Perfect information filtering

- ONLY pass through the task-correlated info → *Discriminability*
- Filter the rest → *Robustness & Generalisation*

Task: Speaker Identification



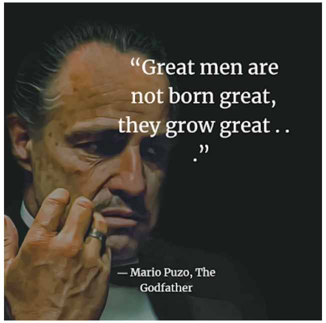
Don Vito Corleone (Marlon Brando)

Perfect Info Processing System (1)

1) Perfect information filtering

- ONLY pass through the task-correlated info → *Discriminability*
- Filter the rest → *Robustness & Generalisation*

Task: Language Identification



English

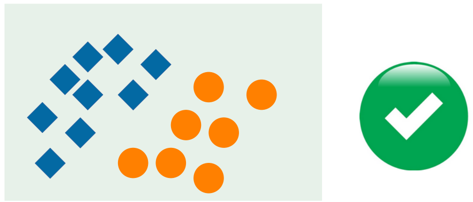
Perfect Info Processing System (2)

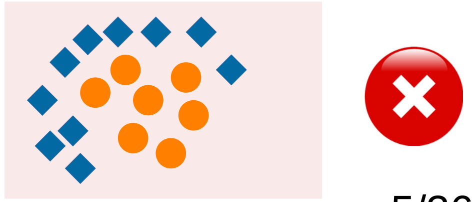
2) Perfect information representation for the classifier

– e.g. Softmax is a linear classifier; likes linearly separable data

What does the classifier expect?

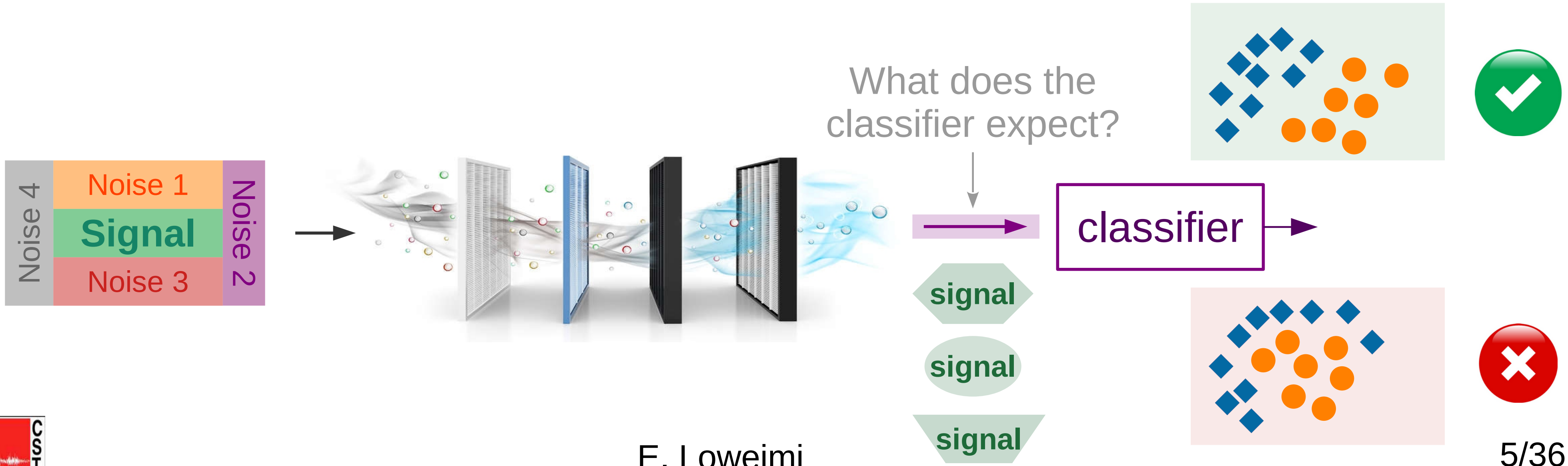
classifier





Perfect Info Processing System (2)

2) Perfect information representation for the classifier
   - e.g. Softmax is a linear classifier; likes linearly separable data



Perfect Info Processing System (3)

3) Input information content

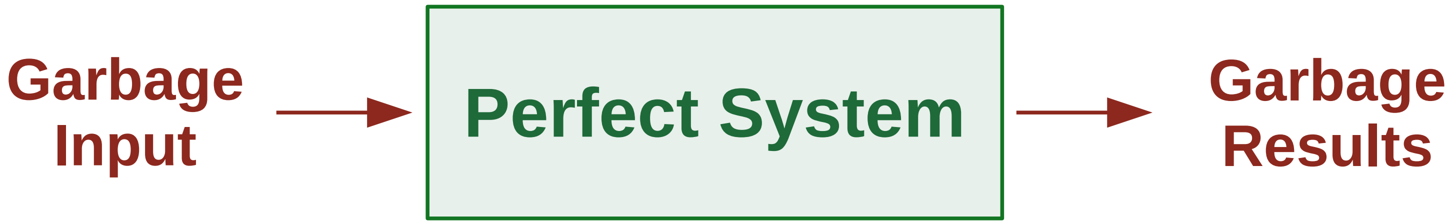
- upper-bounds the effectiveness of info filtering

A perfect info processing system

Perfect Info Processing System (3)

3) Input information content

- upper-bounds the effectiveness of info filtering
- “*Garbage in, Garbage out*”
  - “… *output can only be as accurate as the information entered* ...”

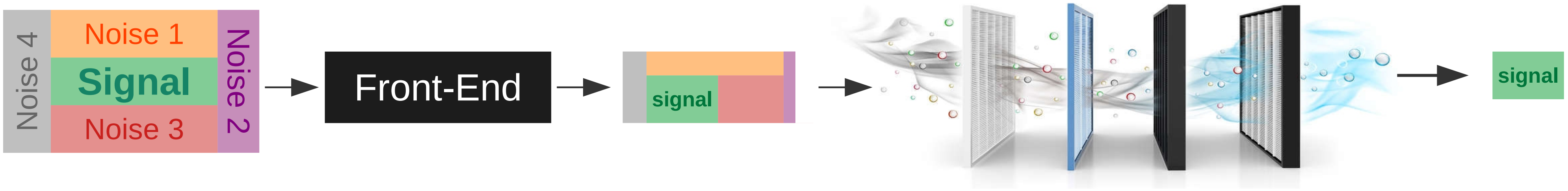


Perfect Info Processing System (3)

3) Input information content

- upper-bounds the effectiveness of info filtering
- "*Garbage in, Garbage out*"
  - "… *output can only be as accurate as the information entered* …"

A perfect info processing system



Perfect Info Processing System (3)

3) Input information content

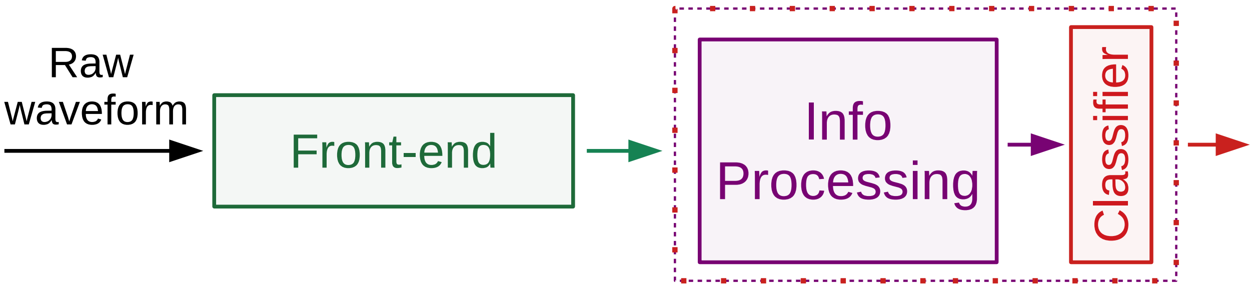
– upper-bounds the effectiveness of info filtering

– *“Garbage in, Garbage out”*

- “… *output can only be as accurate as the information entered* ...”

Our Goal ...

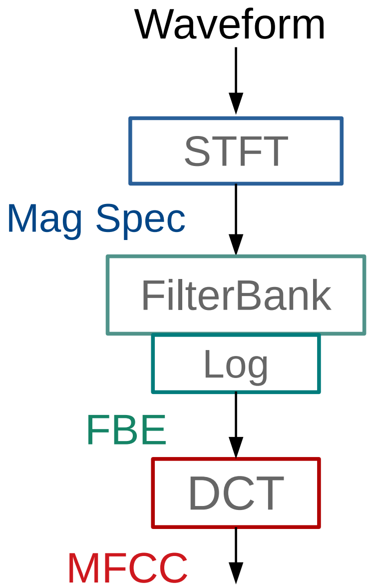
- Components: **Input**, filtering, representation
- **Goal**: Find an input (front-end) that ...
  1) … is as informative as the raw waveform (complete)
  2) … results in a better performance



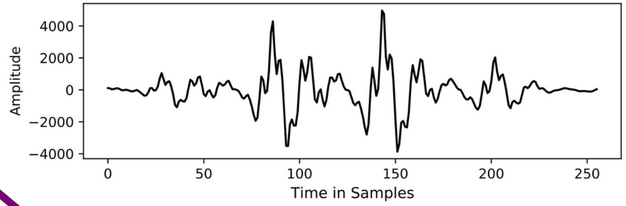
Outline

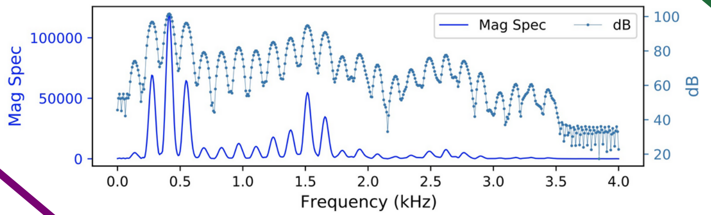
- Motivation
- Signal Information Distribution
- Sign Spectrum
- Combination of Raw Magnitude & Sign Spectra
- Experimental Results
- Conclusion

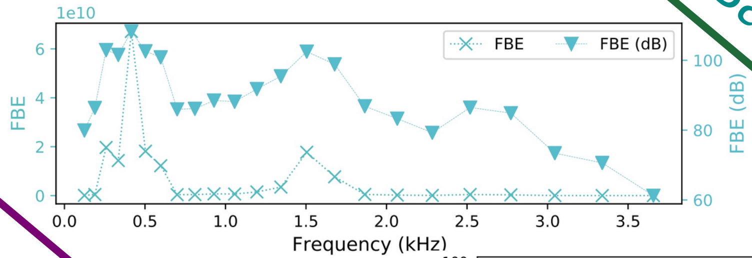
Feature Extraction Pipeline

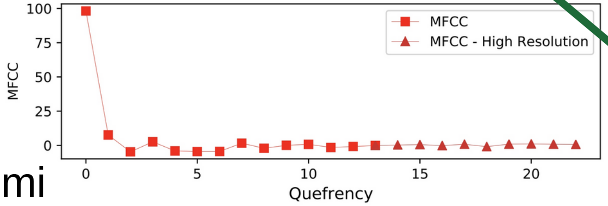


Feature Extraction Pipeline









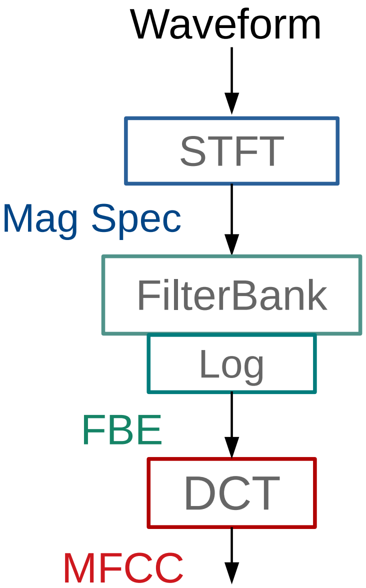
More Information

Raw > Mag > FBank > MFCC

+ More Compact

+ Better Behaviour

− Less Info



Signal Information Distribution (1)

Min-Phase All-pass Decomposition

$$\mathbb{I}_{\text{signal}} = \mathbb{I}_{\text{waveform}} = \mathbb{I}_{\text{Min-Ph}} \cup \mathbb{I}_{\text{All-Pass}}$$

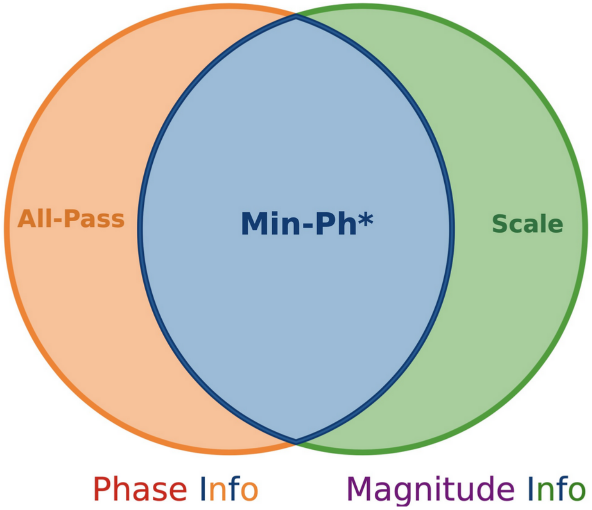
$$\mathbb{I}_{\text{Min-Ph}} = \mathbb{I}_{\text{Scale}} \cup \mathbb{I}_{\text{Min-Ph*}}$$

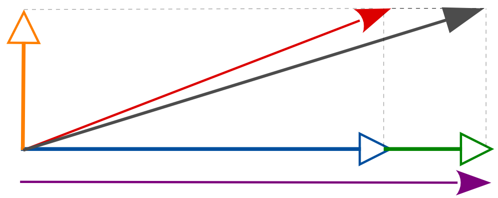
$$\mathbb{I}_{\text{Mag}} = \mathbb{I}_{\text{Min-Ph}}$$

$$\mathbb{I}_{\text{Phase}} = \mathbb{I}_{\text{All-Pass}} \cup \mathbb{I}_{\text{Min-Ph*}}$$

$$\mathbb{I}_{\text{Mag}} \cap \mathbb{I}_{\text{Phase}} = \mathbb{I}_{\text{Min-Ph*}}$$

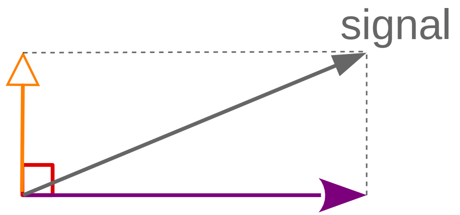
Speech is a **Mixed-phase** signal





All-Pass & Mag in Info Space

- All-pass & Mag spectra are orthogonal in the info space
  - P(AP|Mag) = P(AP)
  - P(Mag|AP) = P(Mag)
  - $\mathbb{I}_{\text{All-Pass}} \cap \mathbb{I}_{\text{Mag}} = \emptyset$



- No chance to recover one from another (underdetermined)
  - No matter how powerful the info processing machinery is!

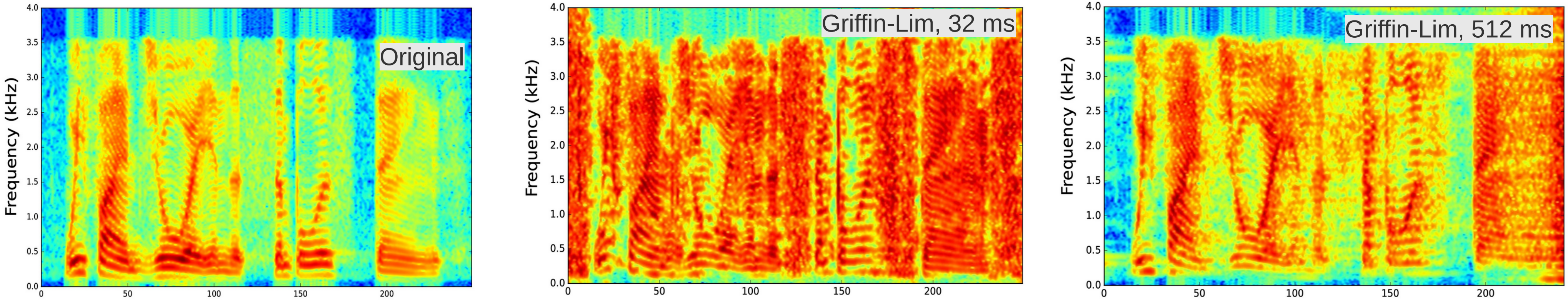
Computing All-Pass Element

- **Parametric** ↔ rational transfer function, H(f)
  - ✗ H(f) may not be available
  - ✗ Finding max-phase zeros is expensive, for a large polynomial

$$\sum_{i=0}^{M} b_i z^i = \prod_{i=1}^{M} (z - z_i)$$

- **Non-parametric** ↔ complex cepstrum
  - ✔ More practical … but involves ...
  - ✗ Phase unwrapping & large FFT size for accuracy

All-Pass Information Content ...



- All-Pass-only reconstructed signal includes ...
  - Temporal localisation info
  - Speech source (excitation) info

* Griffin-Lim (GL)
* overlap: 75%
* #iterations: 100
* window: Hamm

Outline

- Motivation
- Signal Information Distribution
- Sign Spectrum
- Combination of Raw Magnitude & Sign Spectra
- Experimental Results
- Conclusion

Sign Spectrum; An Alternative for All-P ...

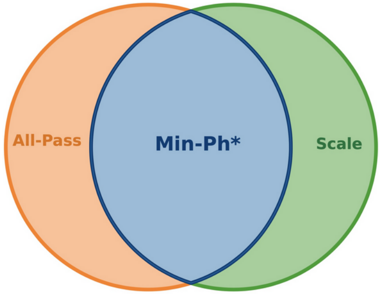
IEEE TRANSACTIONS ON ACOUSTICS, SPEECH, AND SIGNAL PROCESSING, VOL. ASSP-31, NO. 5, OCTOBER 1983

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

- **BOTH** complements magnitude spec info ...

$$\mathrm{I}_{\text{signal}} = \mathbb{I}_{\text{Mag}} \cup \mathbb{I}_{\text{All-Pass}} = \mathbb{I}_{\text{Mag}} \cup \mathbb{I}_{\text{Sign}}$$



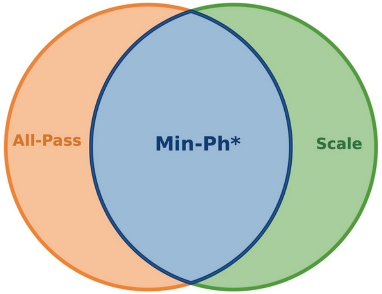
Sign Spectrum; An Alternative for All-P ...

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

- **BOTH** complements magnitude spec info ... **BUT**
  - Sign spectrum samples are ±1 → 1 bit per bin
  - All-Pass samples are float → 16 bits per bin



$$\mathrm{I}_{\text{signal}} = \mathbb{I}_{\text{Mag}} \cup \mathbb{I}_{\text{All-Pass}} = \mathbb{I}_{\text{Mag}} \cup \mathbb{I}_{\text{Sign}}$$

Sign Spectrum - Definition

- One bit of the *phase spectrum* ($\phi_X(\omega)$) info ...

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

$$S_X(\omega;\alpha) = \text{sign}\{\text{Real}\{\exp(j(\frac{\pi}{2} - \alpha)X(\omega))\}\}$$

$$S_X(\omega;\alpha = \frac{\pi}{2}) = \text{sign}\{\text{Real}\{X(\omega)\}\}$$

* $\phi_X(\omega)$ is wrapped (principal) phase → unwrapping is **NOT** needed!

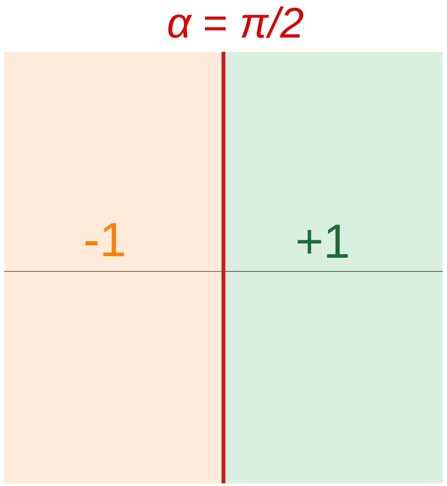
Sign Spectrum - Definition

- One bit of the *phase spectrum* ($\phi_X(\omega)$) info ...

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

$$S_X(\omega;\alpha) = \text{sign}\{\text{Real}\{\exp(j(\frac{\pi}{2} - \alpha)X(\omega))\}\}$$

$$S_X(\omega;\alpha = \frac{\pi}{2}) = \text{sign}\{\text{Real}\{X(\omega)\}\}$$

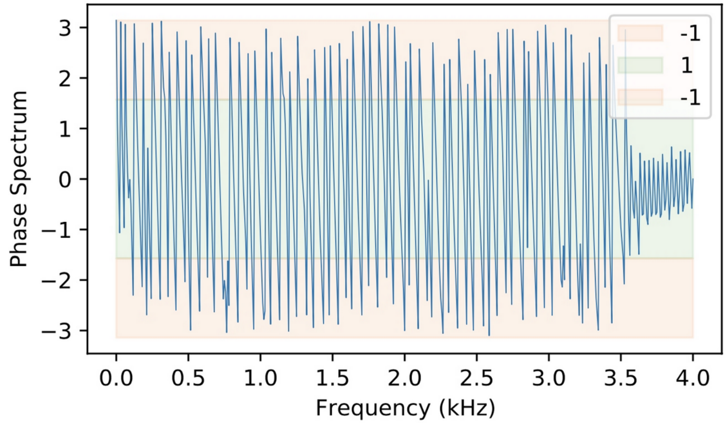


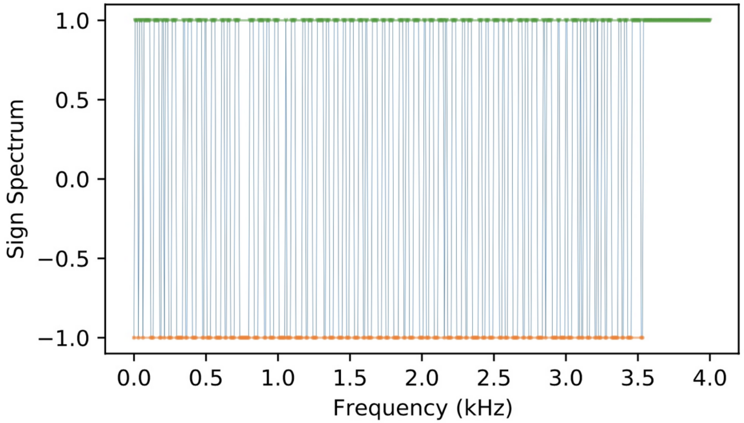
* $\phi_X(\omega)$ is wrapped (principal) phase → unwrapping is **NOT** needed!

Sign Spectrum - Definition

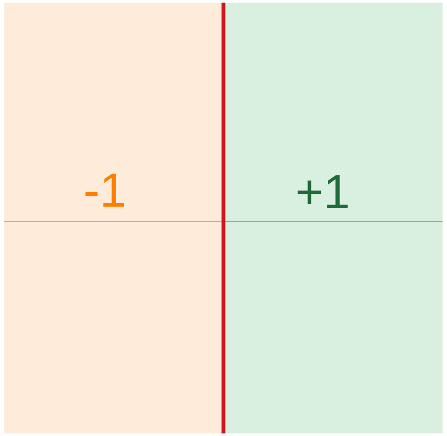
- One bit of the *phase spectrum* ($\phi_x(\omega)$) info ...

$\alpha = \pi/2$





$\alpha = \pi/2$

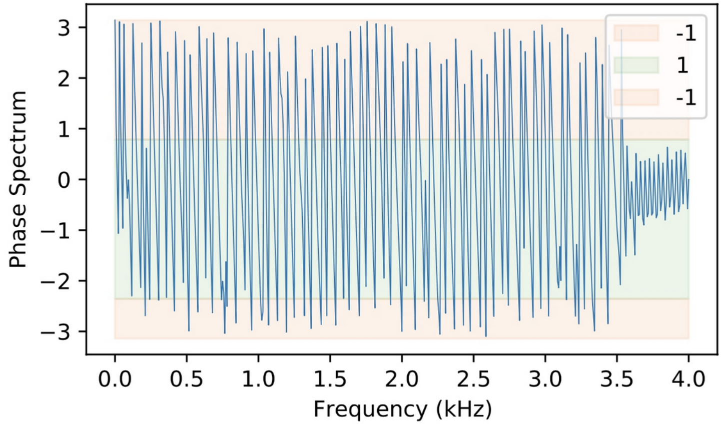


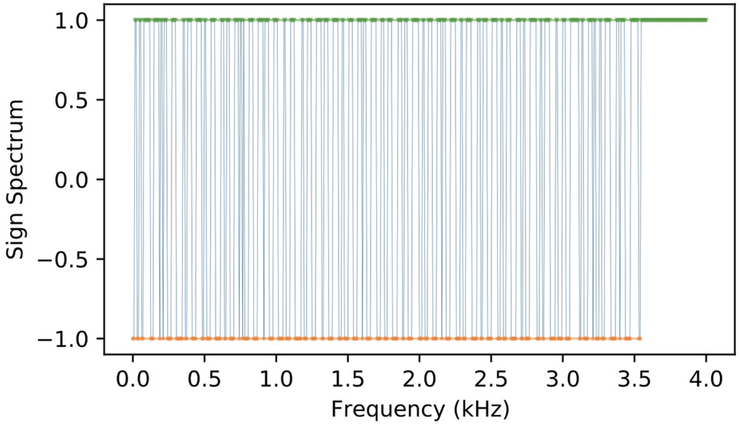
* $\phi_x(\omega)$ is wrapped (principal) phase → unwrapping is **NOT** needed!

Sign Spectrum - Definition

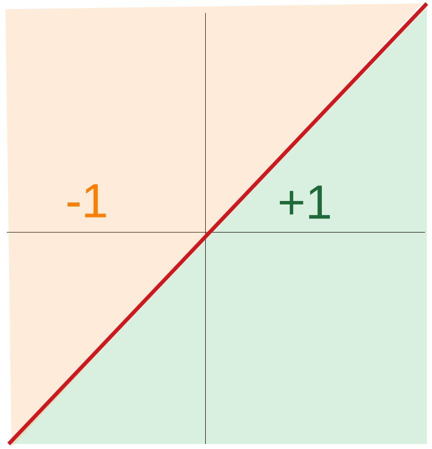
- One bit of the *phase spectrum* ($\phi_x(\omega)$) info ...

$\alpha = \pi/4$



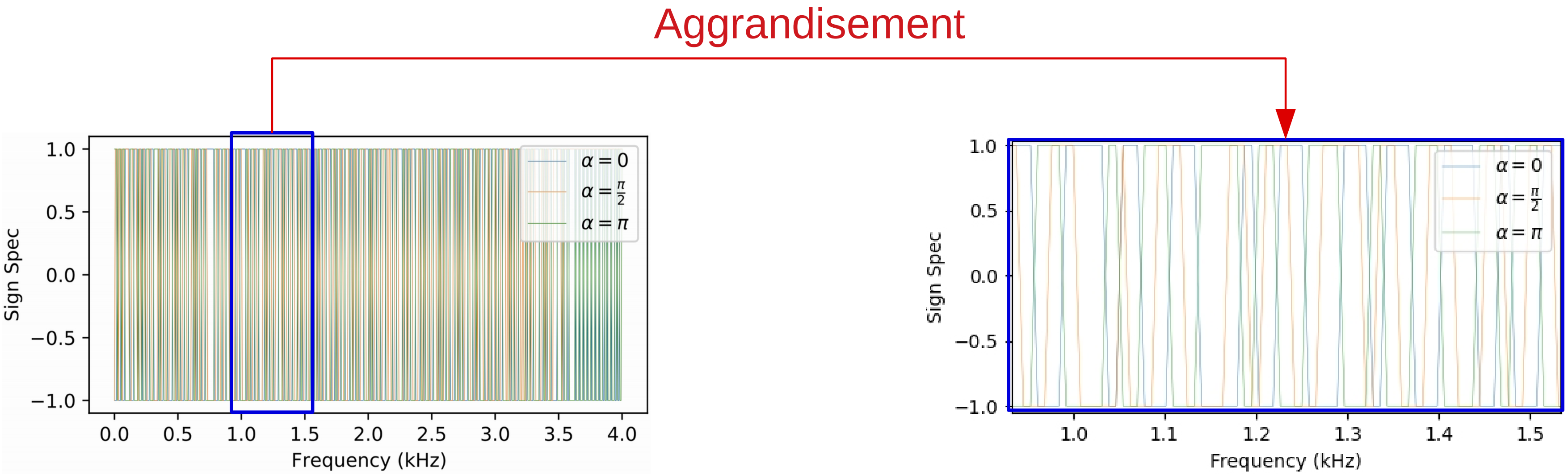


$\alpha = \pi/4$



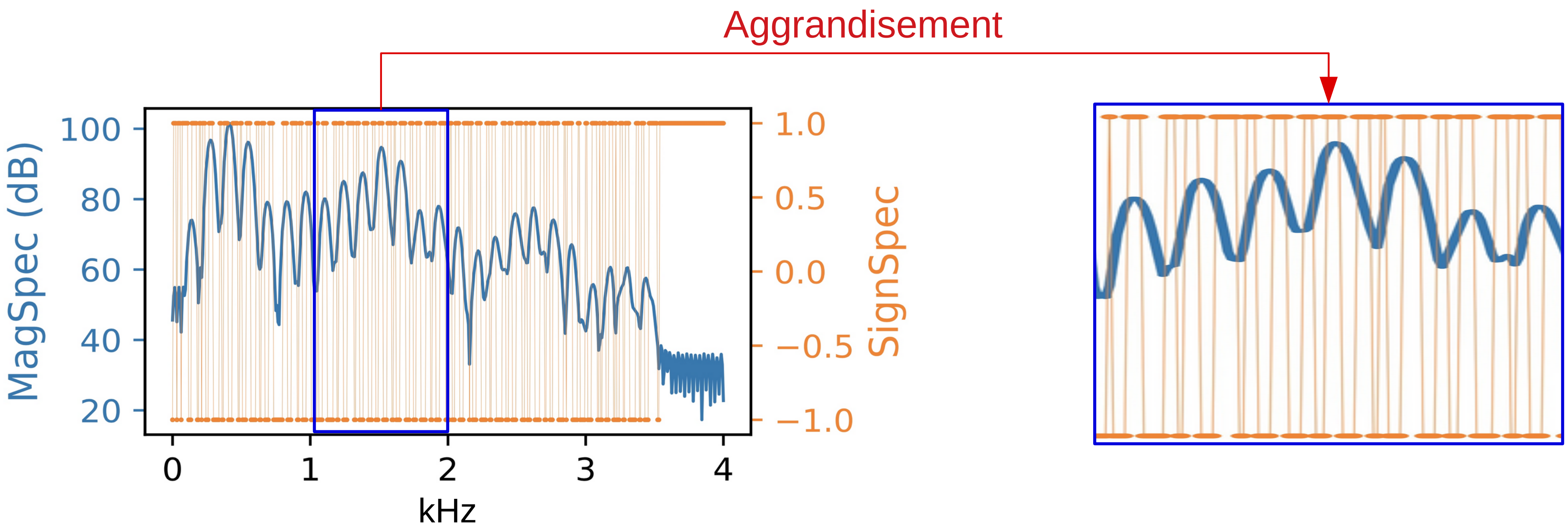
* $\phi_x(\omega)$ is wrapped (principal) phase $\rightarrow$ unwrapping is **NOT** needed!

Understanding Sign Spectrum ...



– sign spectrum is not legible!
– α choice is not important!

Understanding Sign Spectrum ...



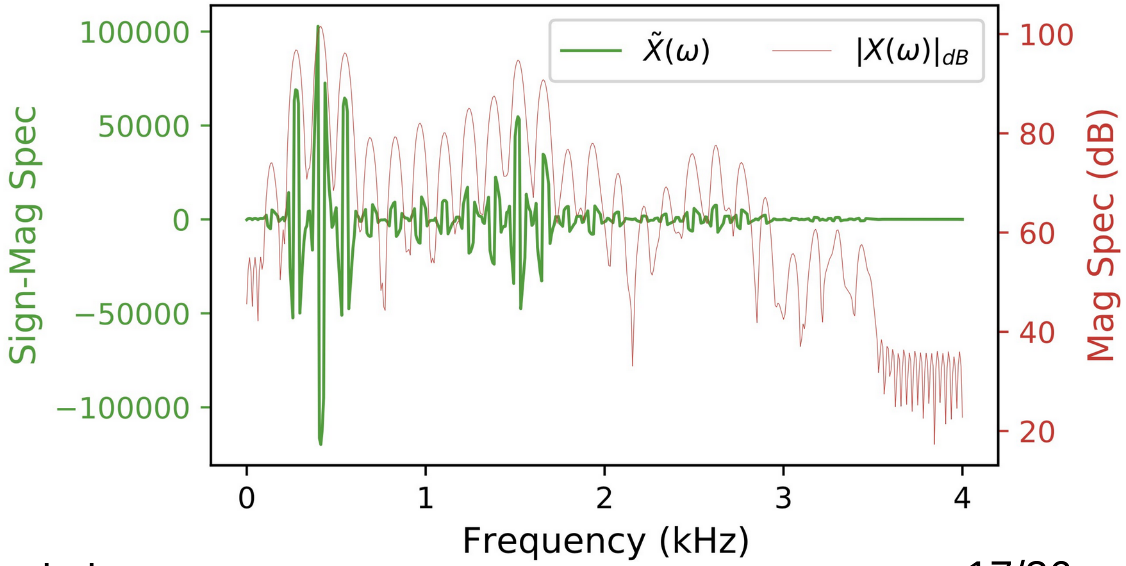
Sign spectrum is not legible!

Signed-Magnitude Spectrum

- Product of the sign and magnitude spectra ...

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

Signed-Magnitude
Spectrum

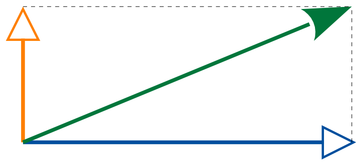


Sign spectrum completes mag info ...

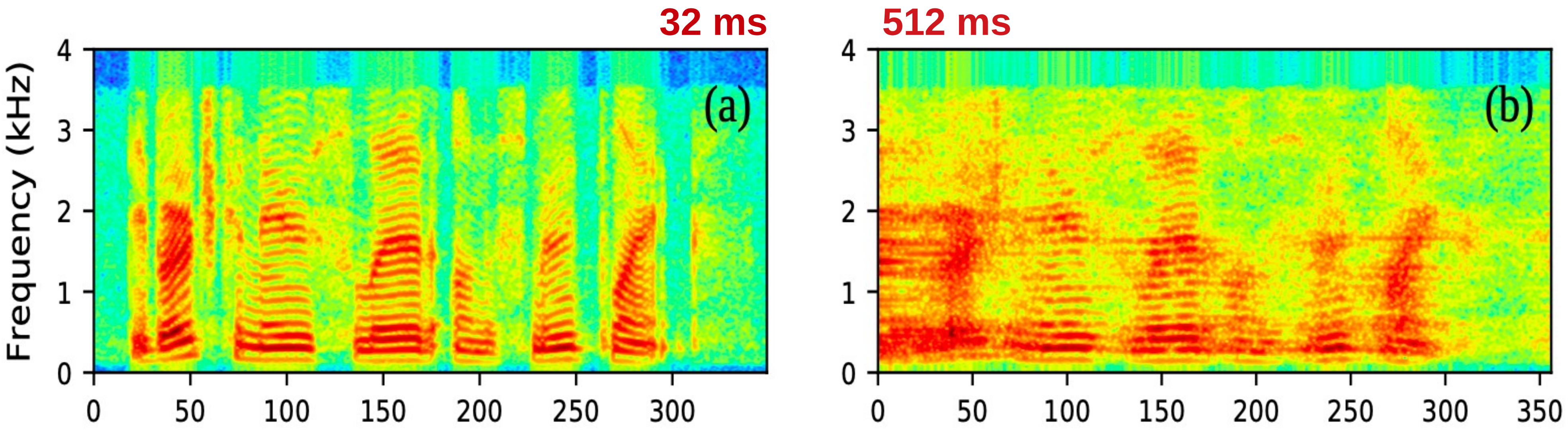
- **Theorem** *Let x[n] and y[n] be two real, causal and finite extent sequence with z-transform which have no zeros on the unit circle. If \tilde{X} = \tilde{Y} for all ω then x[n]=y[n].*

- From information viewpoint …
  - Sign & Mag spectra, together, uniquely characterise *x[n]*

$$\mathbb{I}_{x[n]} = \mathbb{I}_{\tilde{X}(\omega)} = \mathbb{I}_{S_{X(\omega)}} \cup \mathbb{I}_{|X(\omega)|}$$
$$= |\mathbb{I}_{S_{X(\omega)}}| + |\mathbb{I}_{|X(\omega)|}|$$

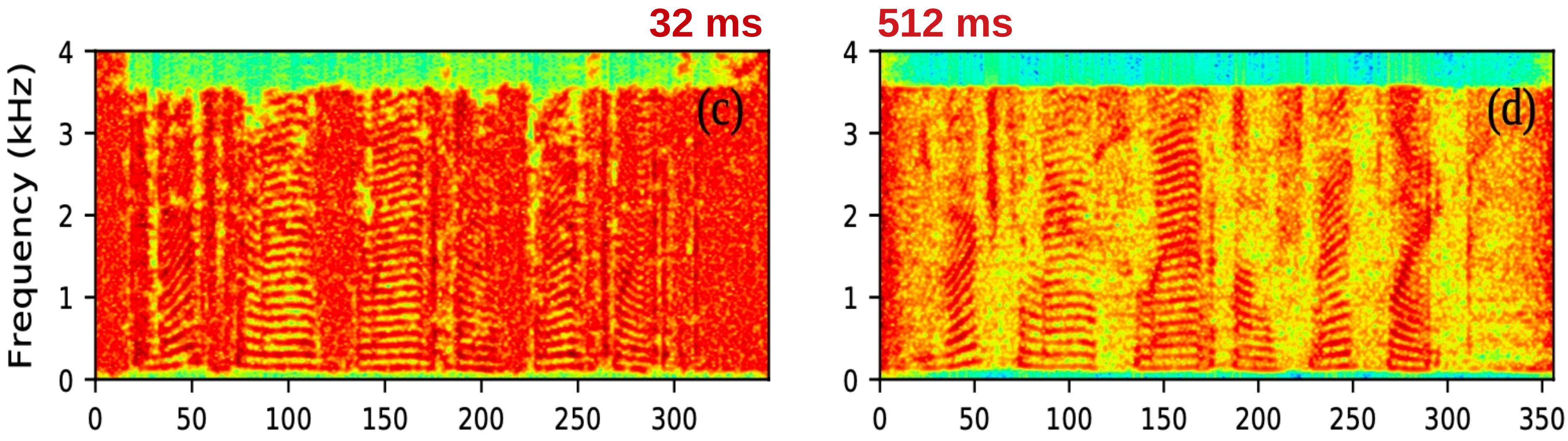


Magnitude-only Signal Reconstruction



* Griffin-Lim (GL) → #iterations: 100; window: Hamming; overlap: 75%
  - PESQ (32 ms): 4.22 ± 0.09
  - PESQ (512 ms): 2.12 ± 0.24

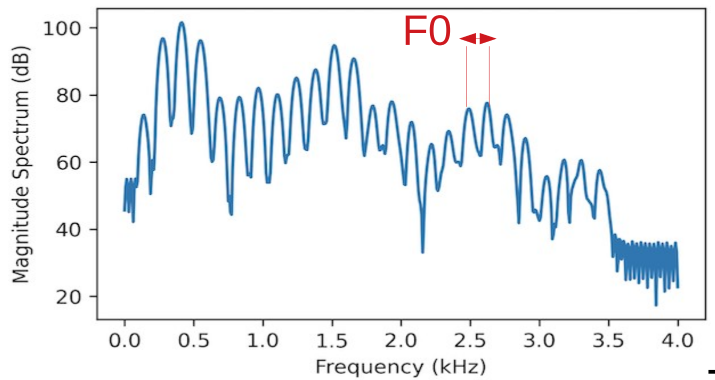
Sign-only Signal Reconstruction via GL

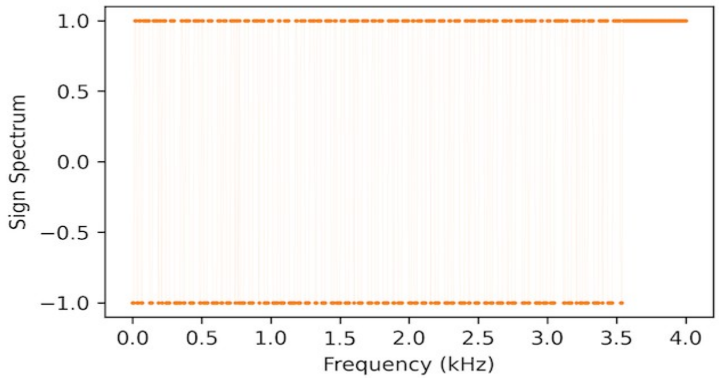


* Sign spectrum info content ...
  – Temporal localisation of events
  – Source (excitation) component info

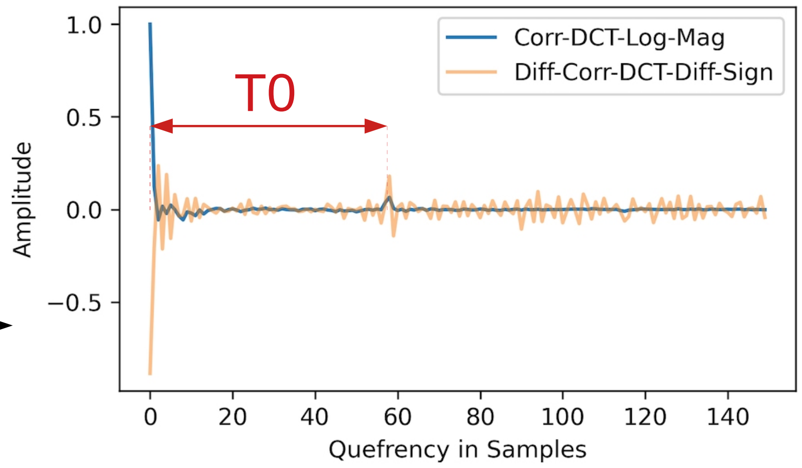
* Griffin-Lim (GL)
* overlap: 75%
* #iterations: 100
* window: Hamm

F0 extraction from Sign Spectrum ...



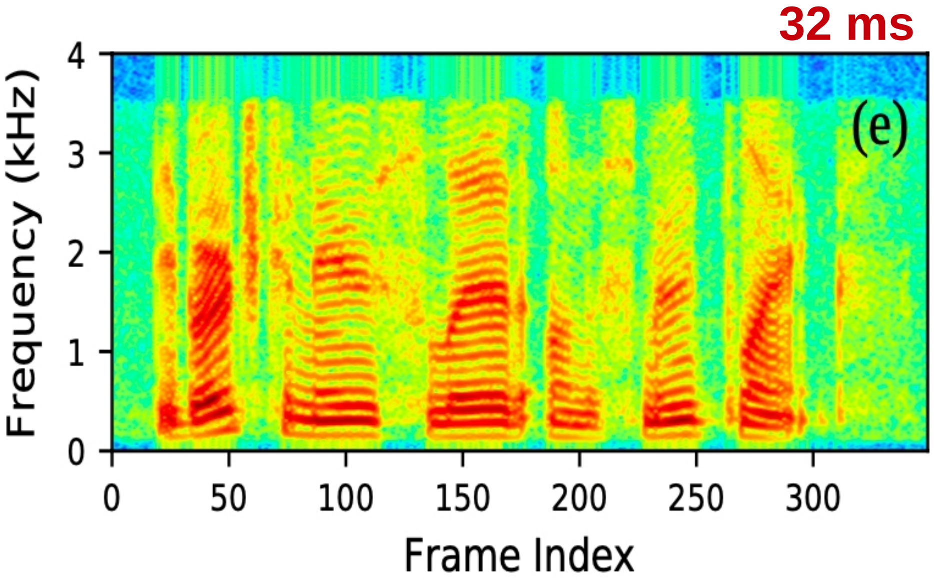


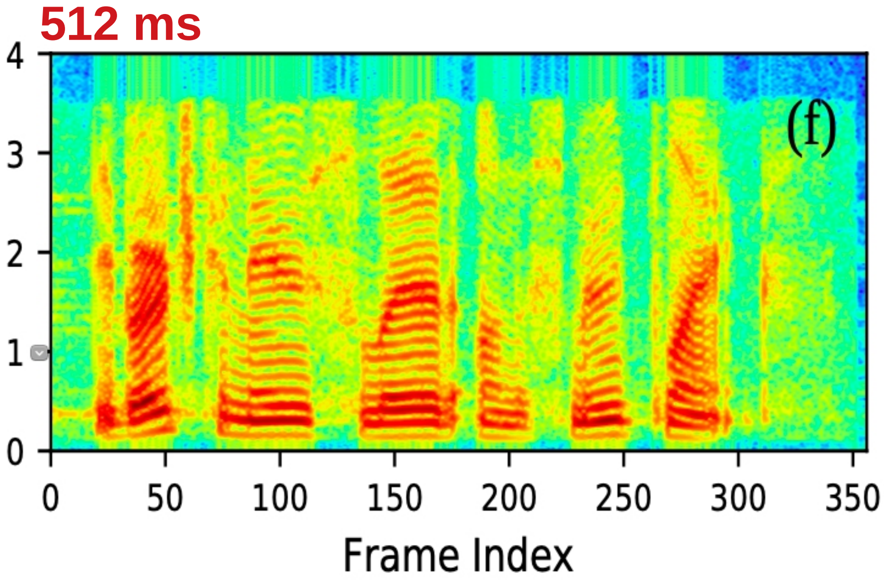
→ F0 extraction →



POSSIBLE! e.g. Sign → Diff → DCT → Corr → Diff

“Mag+Sign”-only Signal Reconstruction





* Griffin-Lim → #iterations: 100; window: Hamming; overlap: 75%

– **NOTE**: Sign spectrum is ONLY used for **initialising** the phase

Playing some <u>mag-only</u> reconstructed signals ...

Original

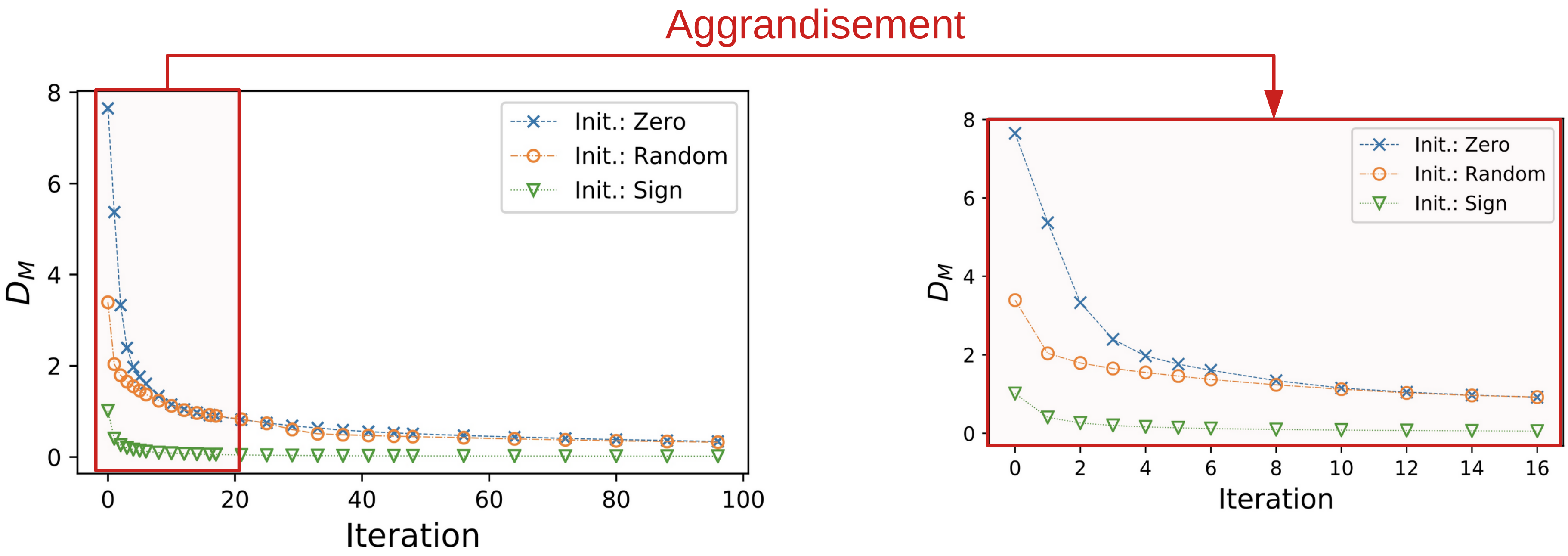
Init. Phase: <u>0</u>

Init. Phase: <u>Random</u>

Init. Phase: <u>Sign</u>

* Griffin-Lim → #iterations: 100; frame length: 32ms; overlap: 75%
* Signal: *sp01.wav* from *NOIZEUS* [sampling rate: 8000 Hz, #bits: 16]
* Text: "*The birch canoe slid on the smooth plank*"

Sign Effect on GL Reconstruction Error



✔ (Near) Perfect reconstruction (error ≈ 0)

✔ Faster convergence

Usefulness of Sign Spec. in PESQ (1)

* Perfect (PESQ=4.5)
* NOT Perfect

	Hamming	
	32 ms	512 ms
Mag	4.22 ± 0.09	2.12 ± 0.24
Mag+Sign	4.50 ± 0.00	4.20 ± 0.08
Gain in PESQ	0.27	2.08

- PESQ (512 ms)[Hamming] $\approx$ 4.2 $\leftarrow$ NOT perfect (4.5)!
  - Does it contradict with the theorem?

Usefulness of Sign Spec. in PESQ (1)

* Perfect (PESQ=4.5)
* NOT Perfect

	Hamming	
	32 ms	512 ms
Mag	4.22 ± 0.09	2.12 ± 0.24
Mag+Sign	4.50 ± 0.00	4.20 ± 0.08
Gain in PESQ	0.27	2.08

- PESQ (512 ms)[Hamming] ≈ 4.2 ← NOT perfect (4.5)!
  - It does **NOT** contradict with the theorem …
  - The theorem tells WHAT is possible, NOT HOW to do it!

Usefulness of Sign Spec. in PESQ (2)

* Perfect (PESQ=4.5)
* NOT Perfect

	Hamming 32 ms	Hamming 512 ms	Rectangular 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

* Griffin-Lim → #iterations: 100; overlap: 75%
  – Gain in PESQ (32 ms) ≈ 0.3
  – Gain in PESQ (512 ms) ≈ 2.1

Outline

- Motivation
- Signal Information Distribution
- Sign Spectrum
- Combination of Raw Magnitude & Sign Spectra
- Experimental Results
- Conclusion

Combination of Sign & Mag for ASR

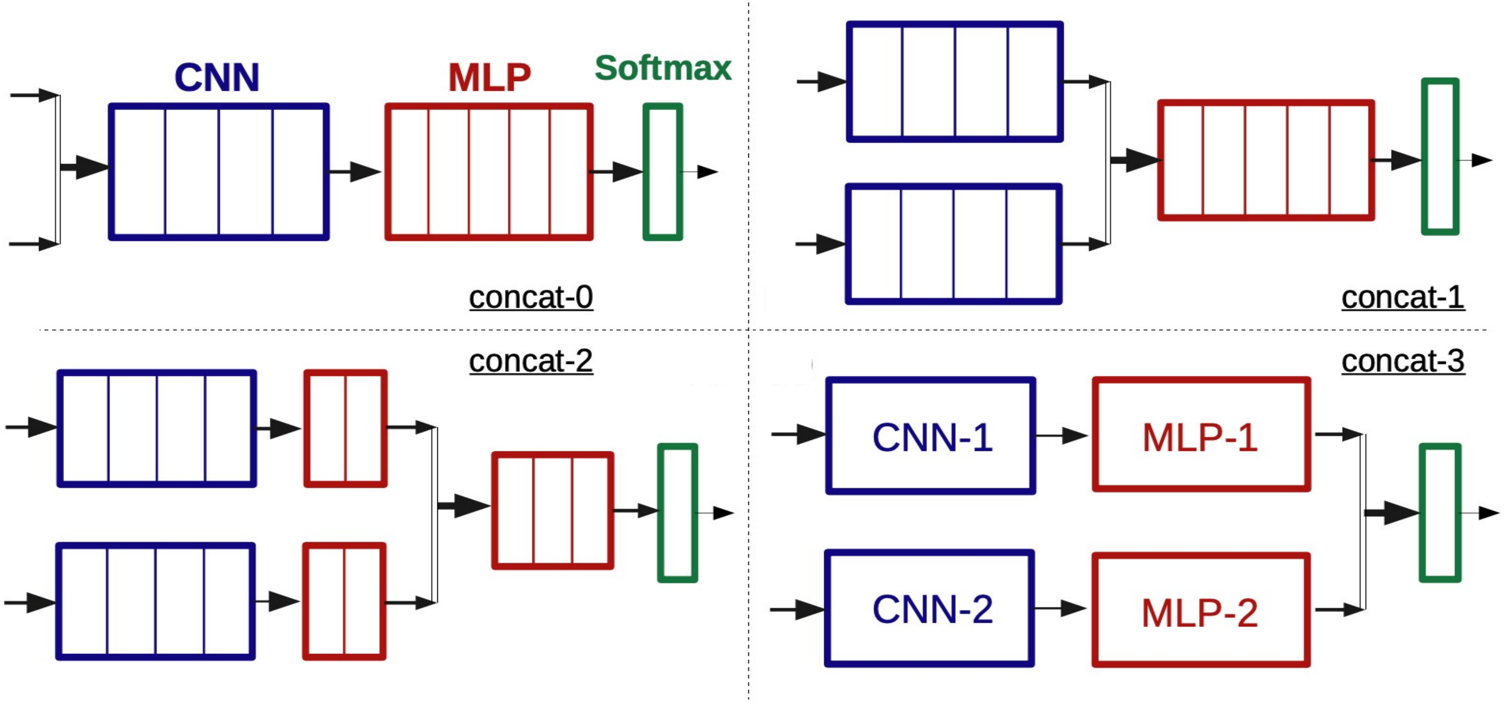
- HOW to combine?
  - In synthesis via FFT/iFFT, e.g. Griffin-Lim → sign times mag
  - In classification/regression → ANYTHING WHICH WORKS!

Combination of Sign & Mag for ASR

- HOW to combine?
  - In synthesis via FFT/iFFT, e.g. Griffin-Lim → sign times mag
  - In classification/regression → ANYTHING WHICH WORKS!

- Multi-stream info processing problem ...
  - How to process each individual stream?
  - How to fuse the (processed) streams?
  - What is the optimal architecture for such task?

Proposed Schemes for Multi-stream Information Processing



CNN
MLP (FC)
Softmax

– What are the pros/cons of each scheme?
– Which one is better? Problem-oriented!

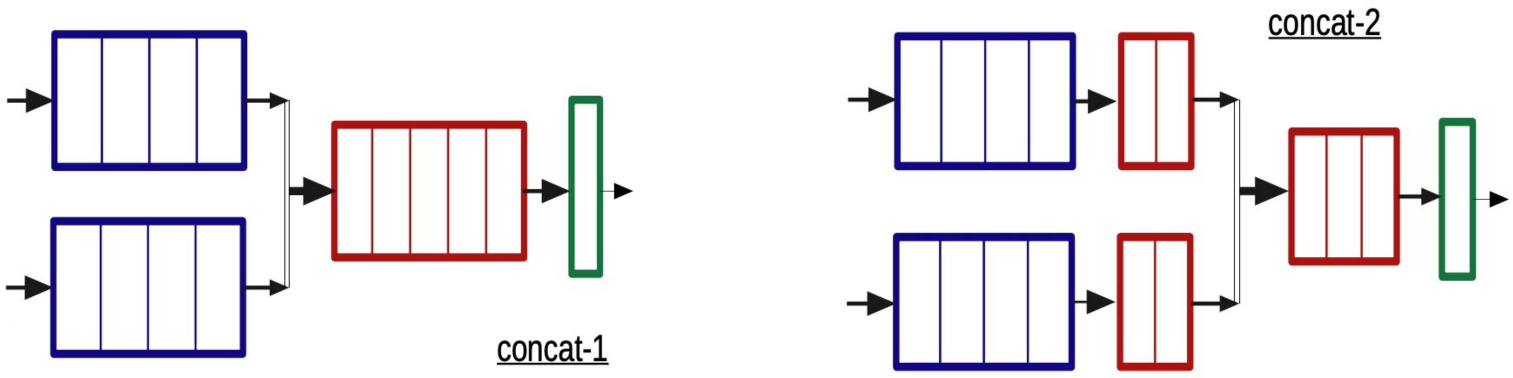
What is the optimal fusion scheme?

- For a given #layers, higher fusion point leads to ...

  1) More layers dedicated to individual stream processing

  - Fewer layers remain for abstraction extraction (after fusion)

CNN
MLP (FC)
Softmax



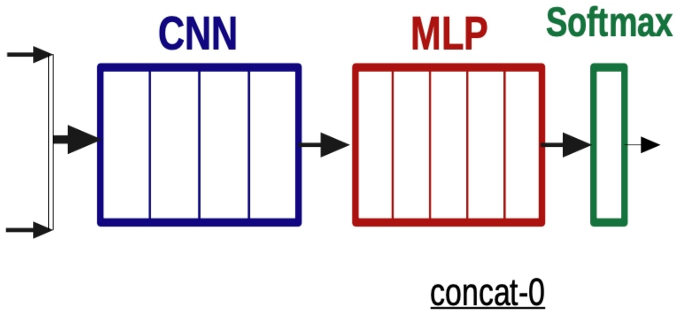
What is the optimal fusion scheme?

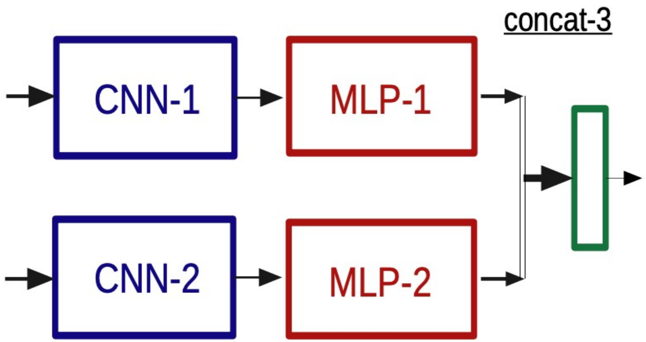
- For a given #layers, higher fusion point leads to …

1) More layers dedicated to individual stream processing

  - Fewer layers remain for abstraction extraction (after fusion)

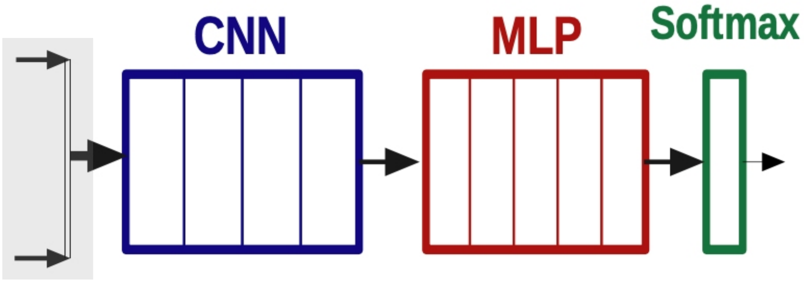
2) More parameters, bigger model, e.g. $\#P_{contant\text{-}3} \approx 2 \times \#P_{concat\text{-}0}$

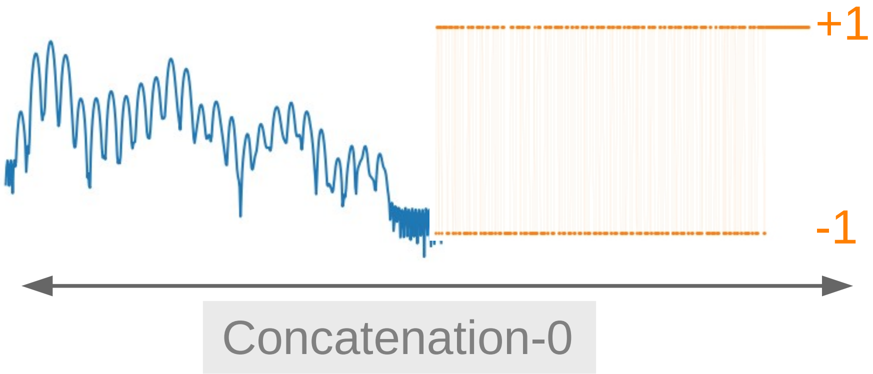




Case Study: Concat-0

- Input Streams: *Mag* and *Sign*
- They r **orthogonal** & differ in ...
  - Info encoding scheme
  - Local patterns/correlation
  - Dynamic range
  - Continuous vs discrete
  - Statistical proprieties

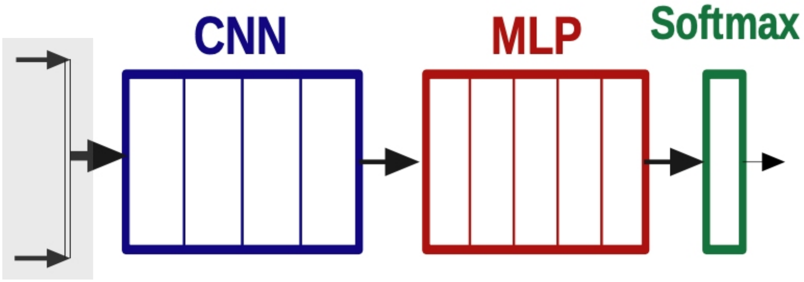


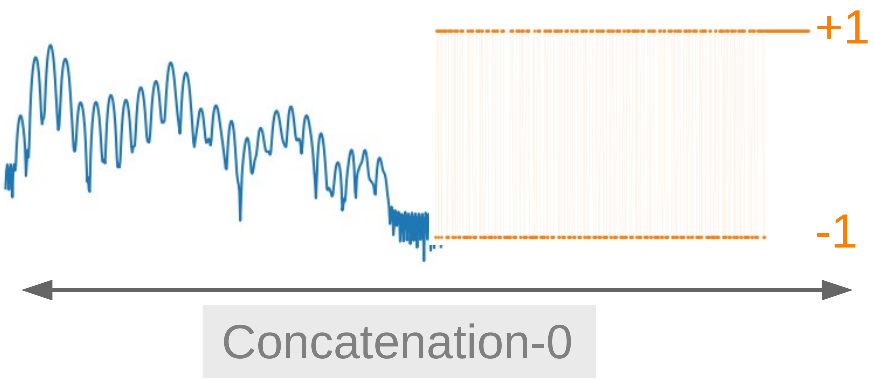


Case Study: Concat-0

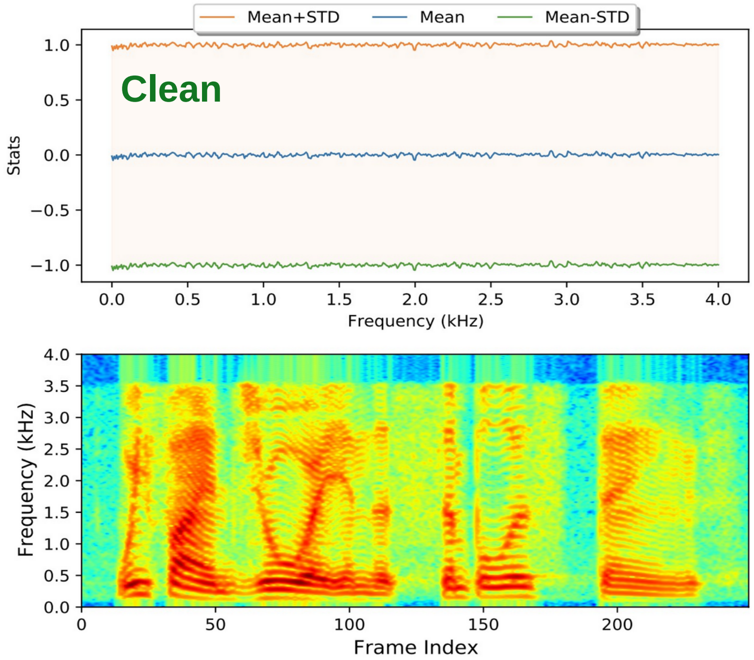
- Input Streams: *Mag* and *Sign*
- They r **orthogonal** & differ in ...
  - Info encoding scheme
  - Local patterns/correlation
  - Dynamic range
  - Continuous vs discrete
  - Statistical proprieties

Using the same set of filters for both perplexes the learner!



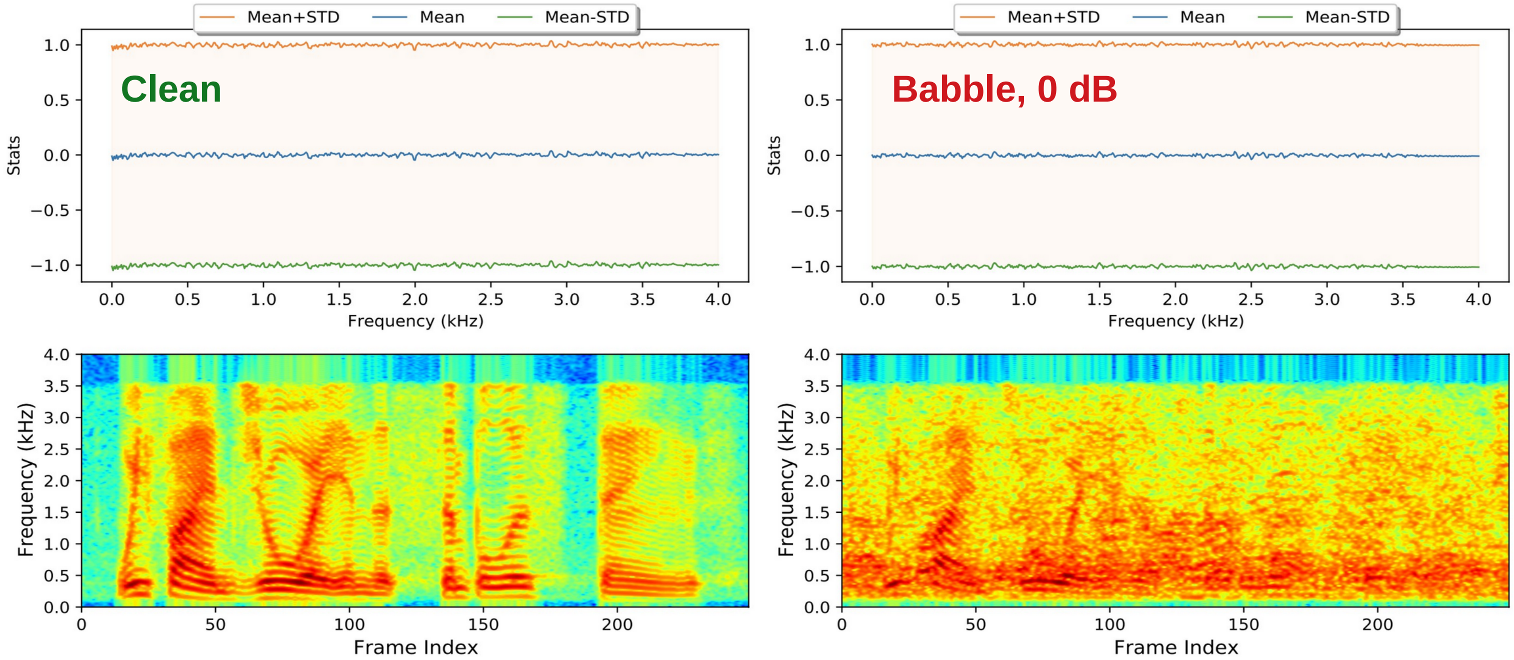


Statistical Properties of Sign Spectrum



- Mean ≈ 0, STD ≈ 1
- Statistical normalisation is not required!

Statistical Properties of Sign Spectrum



- Mean ≈ 0, STD ≈ 1; in all conditions (a structural property)
- Statistical normalisation is not required!

Outline

- Motivation
- Signal Information Distribution
- Sign Spectrum
- Combination of Raw Magnitude & Sign Spectra
- Experimental Results
- Conclusion

Experimental Setup

- Databases: TIMIT/NTIMIT (5.4 h), Aurora-4 (14 h) & WSJ (81 h)
- Toolkit: PyTorch-Kaldi, default setting (w/o monophone regularisation)
- Dropout + {Normalisation: LayerNorm ↔ CNN; BatchNorm ↔ MLP}
- Frame length ± #context_frames:
  - Raw: 200ms ± 0
  - MFCC/FBank/Mag/Sign: 25ms ± 5
- Feature Normalisation: for all features except raw waveform ...
  - Speaker-level MVN for TIMIT/NTIMIT & WSJ
  - Utterance-level MVN for Aurora-4

Experimental Results – TIMIT/NTIMIT

- Mag compression (^0.1) helps
- Sign-only → NOT that bad!
- Mag+Sign concatenation helps
- Mag+Sign is better than Raw
- More info ~~→~~ lower PER

	TIMIT		NTIMIT	
	Dev	Eval	Dev	Eval
MFCC	17.1	18.6	27.5	28.9
FBank	16.3	18.2	27.5	28.5
Raw	17.2	18.6	25.2	26.3
Mag	16.8	17.8	30.9	30.1
$\text{Mag}^{0.1}$	15.9	17.6	25.2	25.6
Sign	27.2	30.0	53.7	54.7
Concat-1	15.4	17.5	24.3	24.8
Concat-2	15.7	17.8	24.8	25.3
Concat-3	15.5	17.5	24.6	25.6

Experimental Results – Aurora-4 (Multi)

- Mag compression helps
- Sign-only → Ave = 31.8%
- Performance …
  - Concat > Mag > FBank > Raw > MFCC >> Sign
  - More info ~~→~~ lower WER
- ✔ Concat-1, Ave-WER = 8.2%

Feature	A	B	C	D	*Ave*
MFCC	3.5	6.8	7.1	16.5	10.7
FBank	2.9	5.9	4.5	14.5	9.2
Raw	3.1	5.7	7.5	16.5	10.3
Mag	2.7	5.5	4.7	14.3	9.0
$\text{Mag}^{0.1}$	2.6	5.3	4.3	14.1	8.8
Sign	7.8	21.5	29.0	46.5	31.8
Concat-1	2.5	5.1	3.9	13.0	8.2
Concat-2	2.4	5.0	4.0	13.6	8.4
Concat-3	2.4	5.1	4.1	13.9	8.6

– A: Clean
– B: Additive Noise
– C: Channel
– B: Additive + Channel

Experimental Results – WSJ

- Mag (^0.1) compression helps
- Sign-only → 21.2%, 14.0%
- Performance ranking …
  - Concat > Raw > Mag > FBank > MFCC
- For WSJ (81 hours)
  - More info → lower WER
  - Concat-1 slightly > 2 & 3

	Dev93	Eval92
MFCC	10.4	6.8
FBank	9.1	5.9
Raw	8.4	5.2
Mag	9.3	5.9
$\text{Mag}^{0.1}$	8.8	5.5
Sign	21.2	14.0
Concat-1	8.1	4.7
Concat-2	8.2	4.8
Concat-3	8.2	4.8

Outline

- Motivation
- Signal Information Distribution
- Sign Spectrum
- Combination of Raw Magnitude & Sign Spectra
- Experimental Results
- Conclusion

Conclusion

- Perfect system ↔ Perfect input → includes all signal info
- Sign spectrum is an alternative for the all-pass component, one bit (±1) of the phase spectrum, completes the magnitude spectrum
- Sign & Magnitude info streams processed via a multi-head CNN
  - Four fusion schemes were investigated
- Notable performance gain was achieved (Aurora-4 & WSJ)
- Future work: Multi-stream "Sign+Mag" processing for other tasks

That's It!

- Thanks for your attention!
- Q&A