SpeechWave

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

[2] King's College London

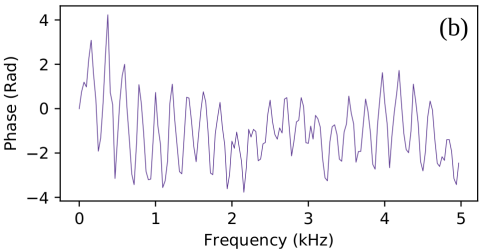
{e.loweimi, peter.bell, s.renals}@ed.ac.uk zoran.cvetkovic@kcl.ac.uk

Outline

- Pros of acoustic modelling using *raw* phase spectrum
- Architectures: Single-stream vs Multi-stream
- Fusion level in multi-stream modelling
- Experimental results
- Conclusion

Acoustic Modelling Using **Raw** Phase Spectrum

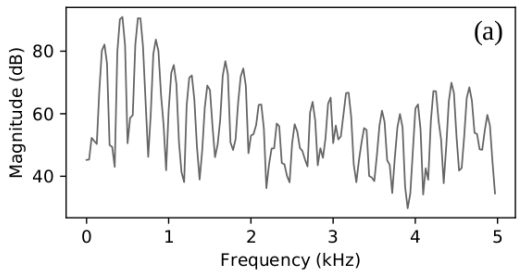
- Raw means using entire spectrum (frequency ≥ 0)
  - If FFT size = 512 → feature size = 257

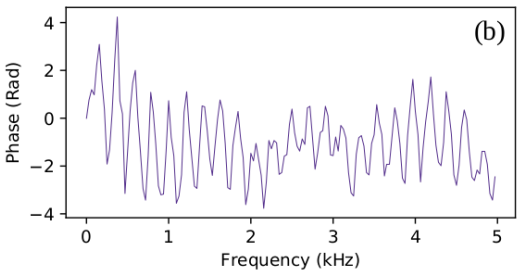


- **Advantages**: Bypass feature engineering
  - Avoid inextricable information loss
  - Dealing with phase's complicated structure

Raw Phase-based Representations

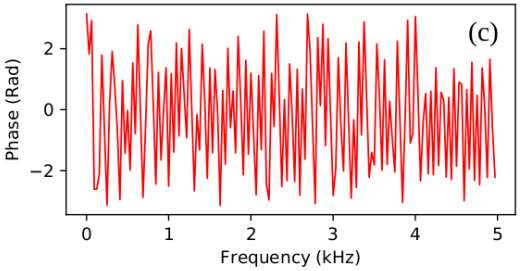
Magnitude Spectrum

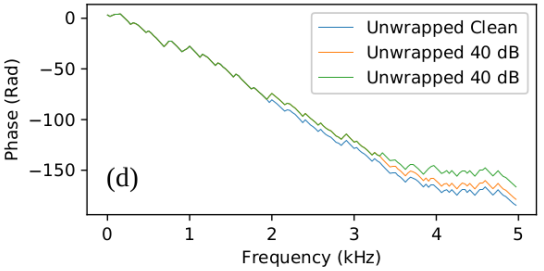




$Phase_{Min\text{-}Ph}$

Wrapped (Principal) Phase

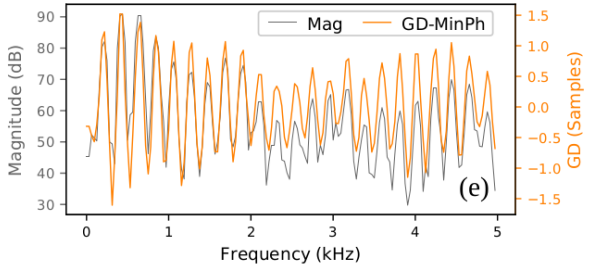


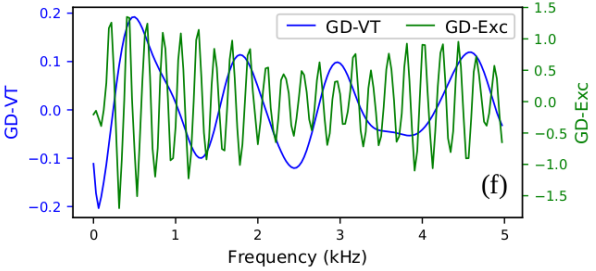


Unwrapped Phase

Mag Spec

$GD_{Min\text{-}Ph}$

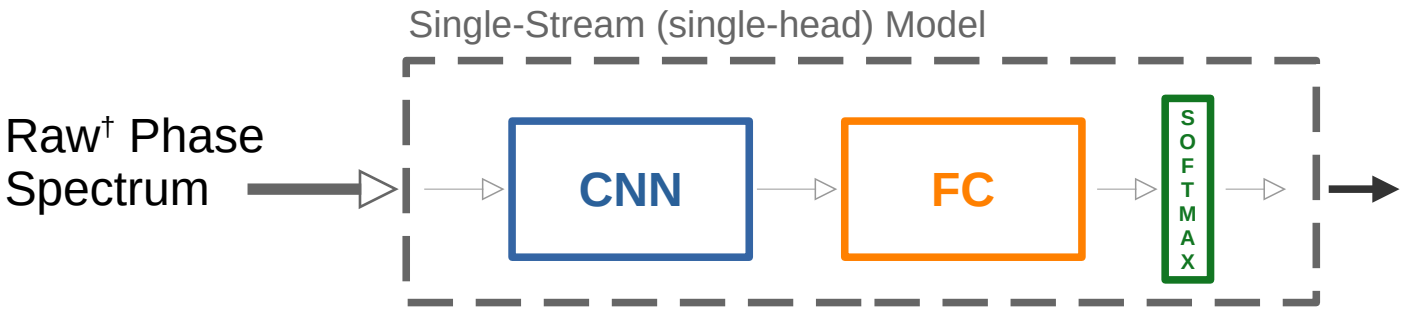




GD_{VT}

GD_{Exc}

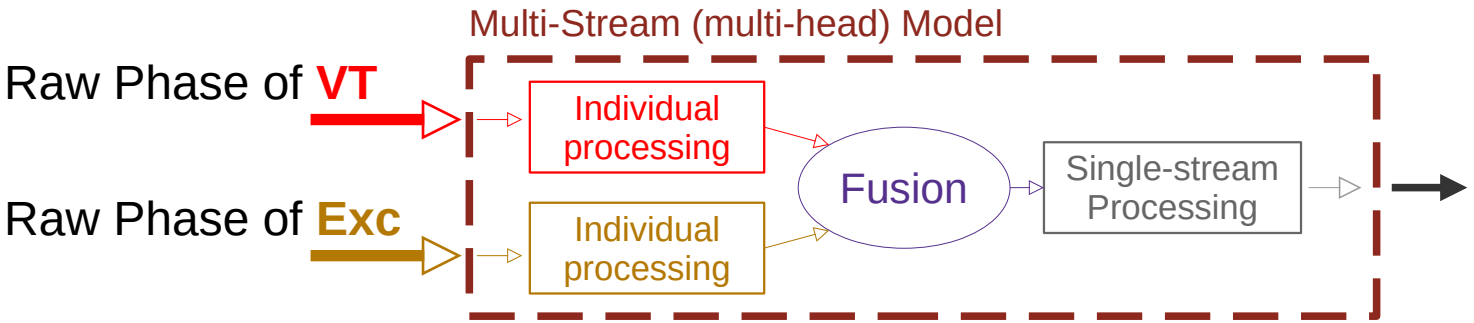
Architecture -- Single-head



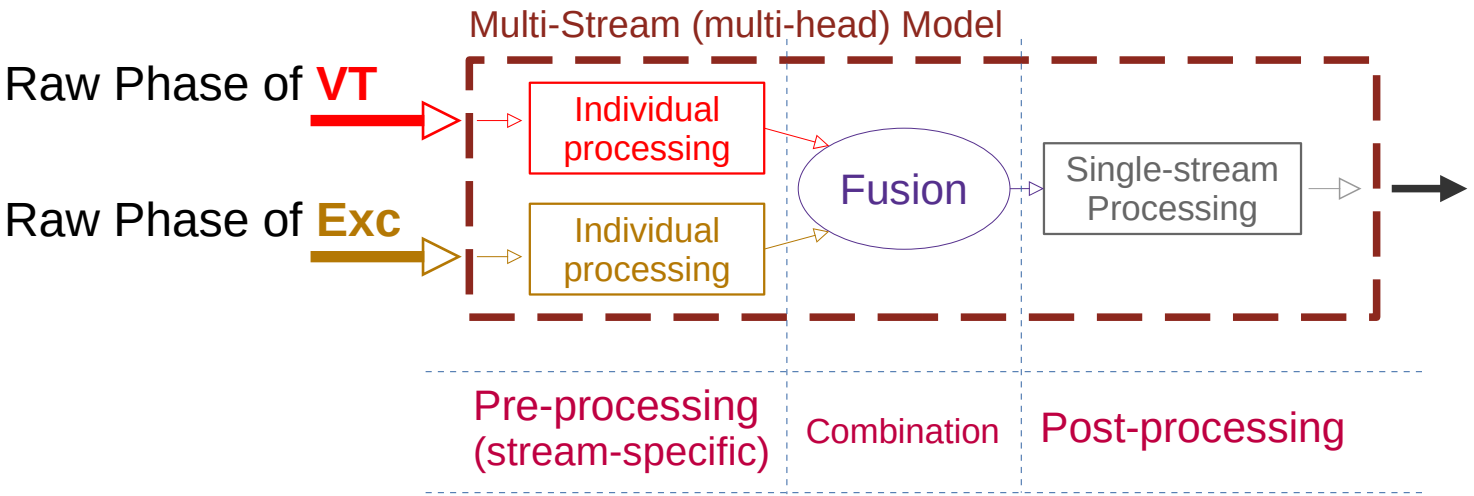
– **Phase-based single stream info**:
  * wrapped, unwrapped, min-phase, source, filter phase spectra ...
  * … or their group delay

– CNN: Convolutional Neural Network
– FC: Fully-connected

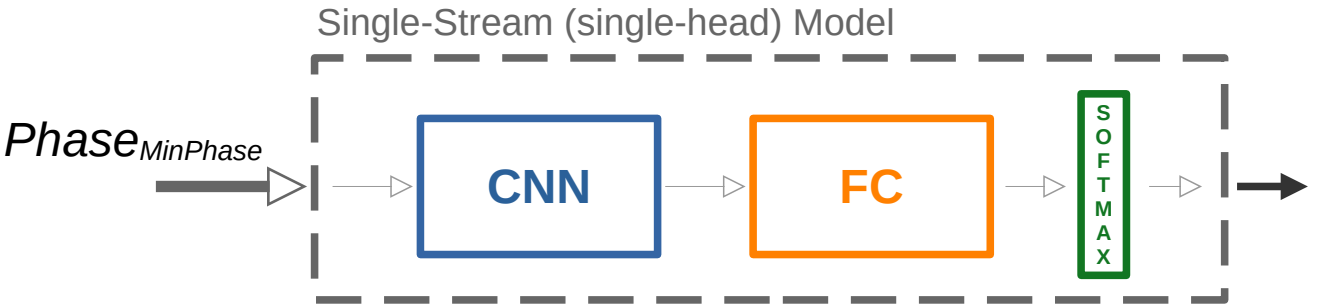
Architecture -- Multi-head

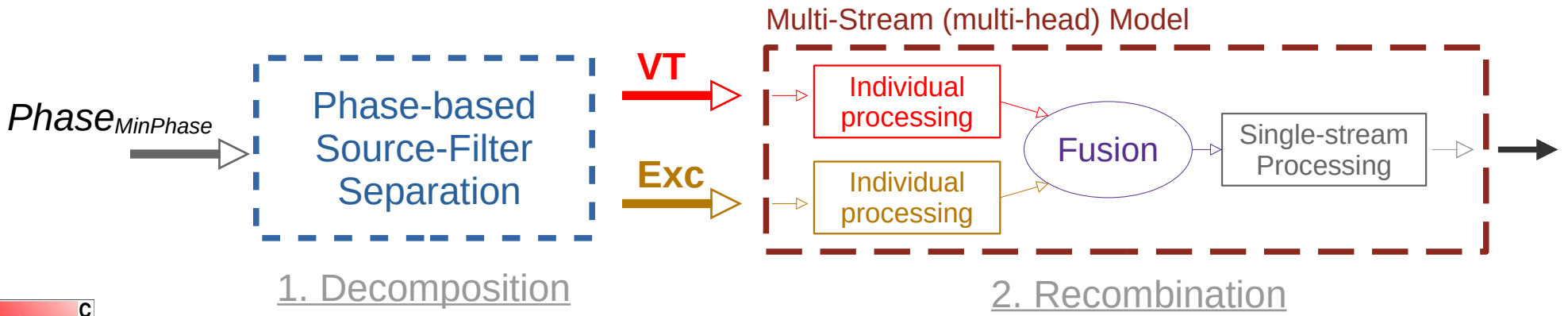


Architecture -- Multi-head

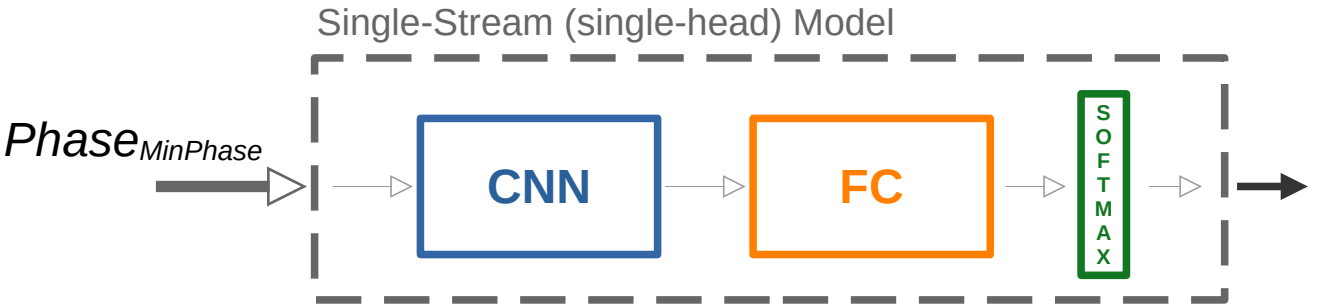


Single-Stream vs Multi-Stream

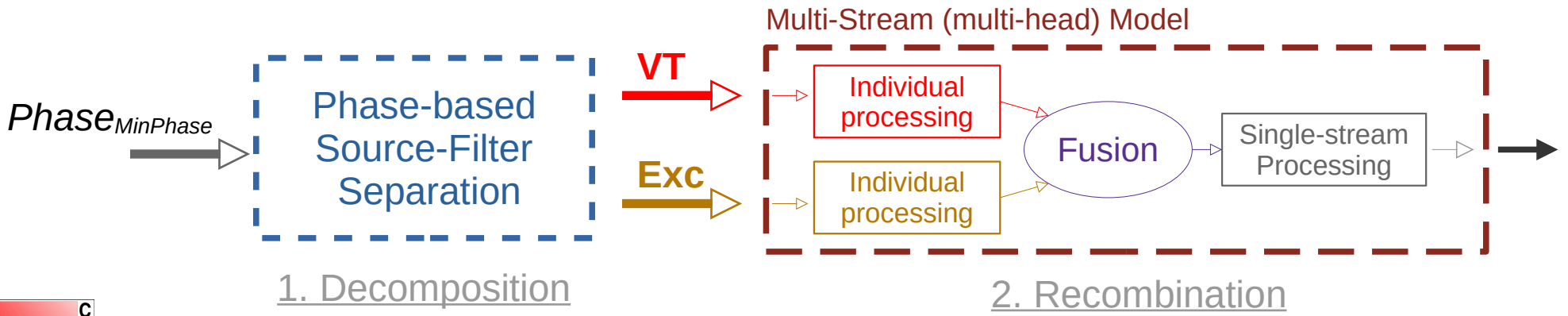




Single-Stream vs Multi-Stream



$$\arg\{X_{MinPh}(\omega)\} = \arg\{X_{VT}(\omega)\} + \arg\{X_{Exc}(\omega)\}$$



Advantages of Multi-Stream: Decomposition-Recombination (1)

- Single-stream
  - Input: $Phase_{MinPhase}$ = **1** x $Phase_{VT}$ **+ 1** x $Phase_{Exc}$
  - Output: $h(Phase_{MinPhase}; \theta_h)$

Advantages of Multi-Stream: Decomposition-Recombination (1)

- Single-stream
  - Input: $Phase_{MinPhase}$ = **1** x $Phase_{VT}$ **+ 1** x $Phase_{Exc}$
  - Output: $h(Phase_{MinPhase}; \theta_h)$
- Multi-stream
  - Input: $Phase_{VT}$ & $Phase_{Exc}$ info streams
  - Output: $h([f(VT; \theta_f) , g(Exc; \theta_g)]; \theta_h)$

concatenation

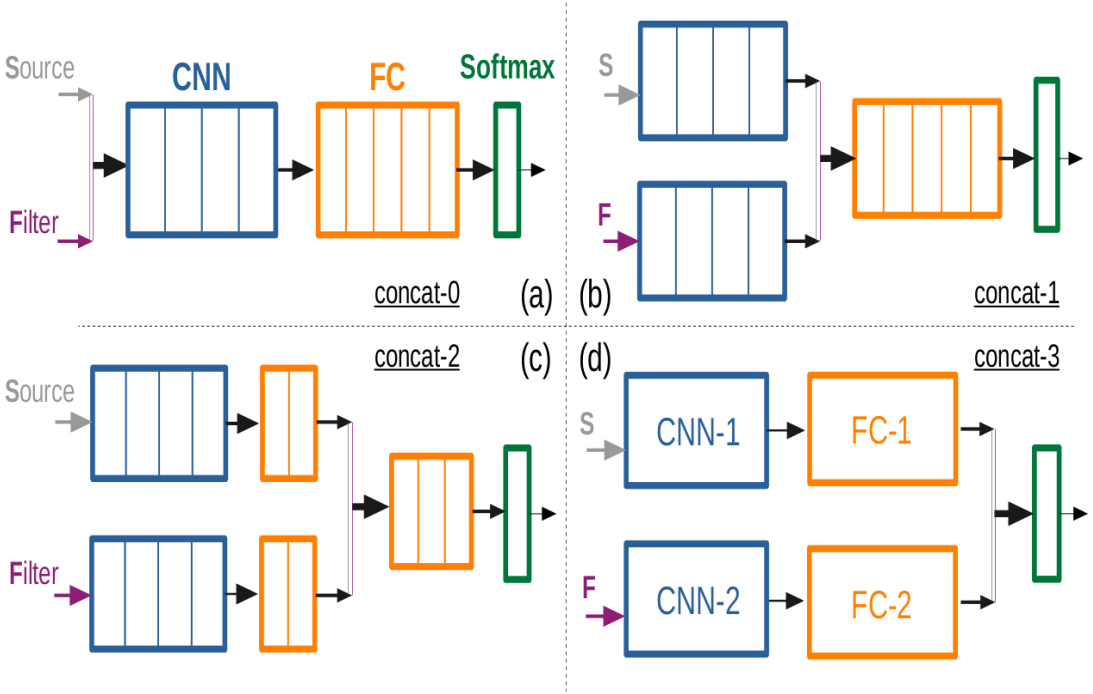
Advantages of Multi-Stream: Decomposition-Recombination (2)

- Each stream (VT & Exc) is weighted/gated properly
  - ... importance to the task
- Learning bespoke transforms for each stream
  - Optimal chain of transforms for each stream is different
- Information fusion at optimal level of abstraction
  - … instead of additive fusion at input level, $\text{Phase}_{VT} + \text{Phase}_{Exc}$

Multi-Stream Architecture

- Fusion@HigherLevels
  - More parameters
  - More pre-processing
  - Less post-processing

- Optimal Trade-off ???



Fusion@HigherLevels: Concat-0 → … → Concat-3

Experimental Setup

- Models built using PyTorch-Kaldi
  - CNNs: 4 layers, 1D, LayerNorm, ReLU
  - FCs: 5 Layers, BatchNorm, ReLU
- Alignments: Kaldi
- Tasks/Measure: TIMIT/PER and WSJ/WER
- Phase-based Source-Filter Separation based on [20]
- MVN@SpeakerLevel; Append: ±5 context frames
- No data augmentation or rescoring with RNNLM

Experimental Results

TIMIT – Phone Recognition

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
$\text{Mag}^{0.1}$	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

WSJ – LVCSR

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
Mag	9.3	5.9	9.1
$\text{Mag}^{0.1}$	8.8	5.5	9.0
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	8.3	5.1	7.8
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	7.9	4.8	7.4
Concat-2	8.1	4.8	7.7
Concat-3	8.2	5.0	8.1

Acoustic modelling using raw phase spectrum works ...

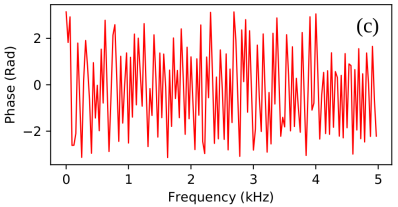
Discussion (1)

- Compared with magnitude-based features, comparable to better WERs are achieved using raw phase spectrum.

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
Mag	9.3	5.9	9.1
$\text{Mag}^{0.1}$	8.8	5.5	9.0
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	8.3	5.1	7.8
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	7.9	4.8	7.4
Concat-2	8.1	4.8	7.7
Concat-3	8.2	5.0	8.1

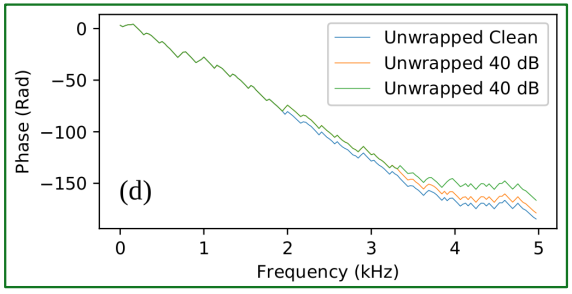
Discussion (2)



- Decent WER for wrapped phase
- Unwrapping increases WER
  - Instability ...
- Using GD improves WER

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
Mag	9.3	5.9	9.1
$\text{Mag}^{0.1}$	8.8	5.5	9.0
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	8.3	5.1	7.8
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	7.9	4.8	7.4
Concat-2	8.1	4.8	7.7
Concat-3	8.2	5.0	8.1



Discussion (3)

- Multi-stream outperforms single-stream
  - ... NO EXTRA INFO in multi-stream ...
- Optimal fusion level ↔ Concat-1
  - Trade-off between ...

    pre- & post-processing

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
Mag	9.3	5.9	9.1
$\text{Mag}^{0.1}$	8.8	5.5	9.0
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	8.3	5.1	7.8
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	7.9	4.8	7.4
Concat-2	8.1	4.8	7.7
Concat-3	8.2	5.0	8.1

Conclusion

- **Goal**: Acoustic modelling using speech's raw phase spectrum
- **Architectures**:
  - Single-head ← raw $\text{Phase}_{\text{wrapped}}$, $\text{Ph}_{\text{unwrapped}}$, Ph_{MinPh}, Ph_{VT}, Ph_{Exc}
  - Multi-head/stream ← raw Source and Filter phase spectra
- **Advantages** of multi-stream approach & optimal fusion level discussed
- **Tasks**: Phone recognition (TIMIT), LVCSR (WSJ)
- **Future Work**: the proposed multi-stream phase-based approach is a general framework, potentially applicable to a wide range of tasks

That's it!

- Thanks for your attention!
- Q & A

- Appendices:
  - Source-filter separation in the phase domain
  - Group delay (GD)

SpeechWave

Phase-based Source-Filter Separation (1)

$$x[n] = x_{VT}[n] * x_{Exc}[n]$$

$$\log|X(\omega)| = \log|X_{VT}(\omega)| + \log|X_{Exc}(\omega)|$$

Hilbert Trans.

$$\arg\{X_{MinPh}(\omega)\} = -\frac{1}{2\pi}\log|X(\omega)| * \cot(\frac{\omega}{2})$$

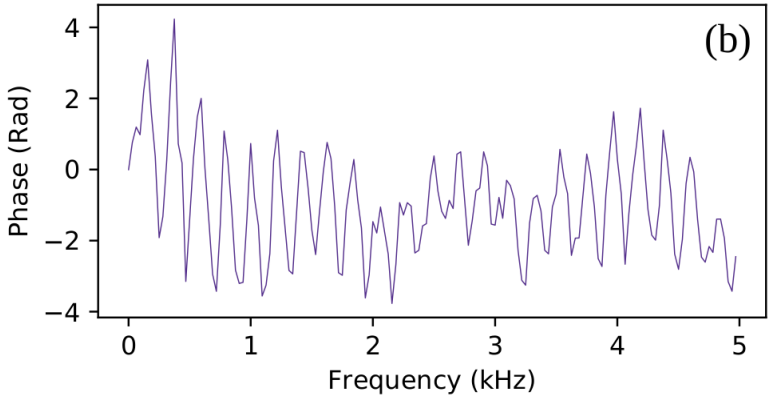
$$\arg\{X_{MinPh}(\omega)\} = \arg\{X_{VT}(\omega)\} + \arg\{X_{Exc}(\omega)\}$$

$$\mathrm{GD}_{MinPh}(\omega) = \mathrm{GD}_{VT}(\omega) + \mathrm{GD}_{Exc}(\omega)$$

Source and Filter components are **additive** in the Log-Mag, Min-phase phase or group delay domains.

Phase-based Source-Filter Separation (2)

$$\arg\{X_{MinPh}(\omega)\} = \arg\{X_{VT}(\omega)\} + \arg\{X_{Exc}(\omega)\}$$

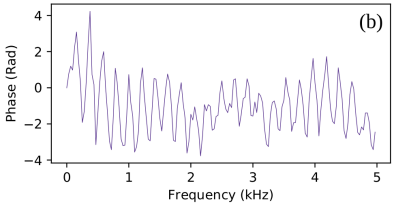


Phase$_{Min\text{-}Ph}$

Phase$_{Min\text{-}Ph}$ ≡ Trend + Fluctuation ≡ Filter (VT) + Source (Exc)

Phase-based Source-Filter Separation (3)

$$\arg\{X_{MinPh}(\omega)\} = \arg\{X_{VT}(\omega)\} + \arg\{X_{Exc}(\omega)\}$$



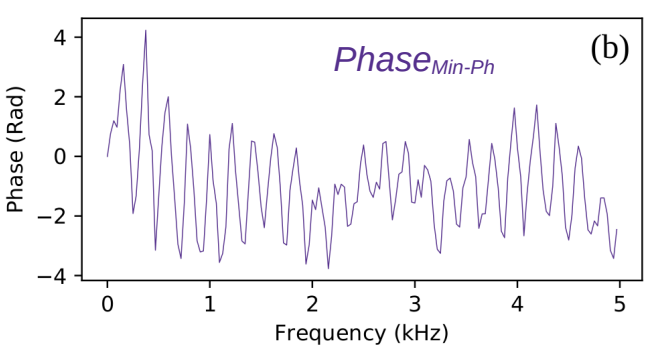
$$arg\{X_{VT}(\omega)\} \approx arg\{X_{MinPh}(\omega)\} * h_{LPF}(\omega)$$

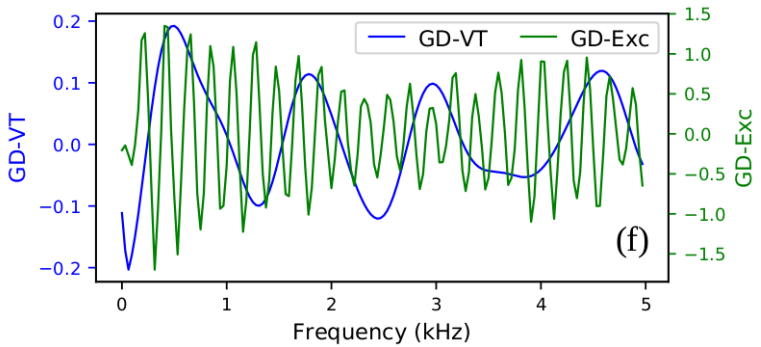
$$arg\{X_{Exc}(\omega)\} = arg\{X_{MinPh}(\omega)\} - arg\{X_{VT}(\omega)\}$$

Phase$_{Min\text{-}Ph}$ ≡ Trend + Fluctuation ≡ Filter (VT) + Source (Exc)

Phase-based Source-Filter Separation (3)

$$\arg\{X_{MinPh}(\omega)\} = \arg\{X_{VT}(\omega)\} + \arg\{X_{Exc}(\omega)\}$$

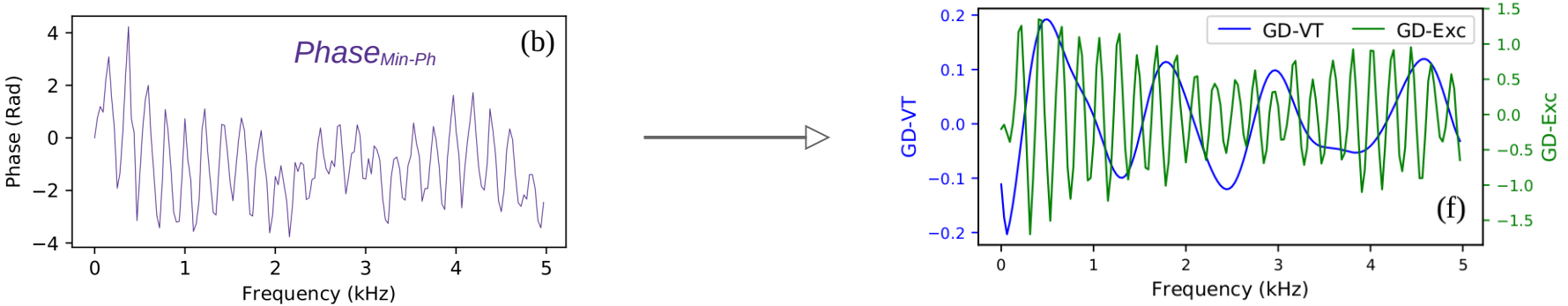




Phase$_{Min\text{-}Ph}$ ≡ Trend + Fluctuation ≡ Filter (VT) + Source (Exc)

Phase-based Source-Filter Separation (3)

$$\arg\{X_{MinPh}(\omega)\} = \arg\{X_{VT}(\omega)\} + \arg\{X_{Exc}(\omega)\}$$

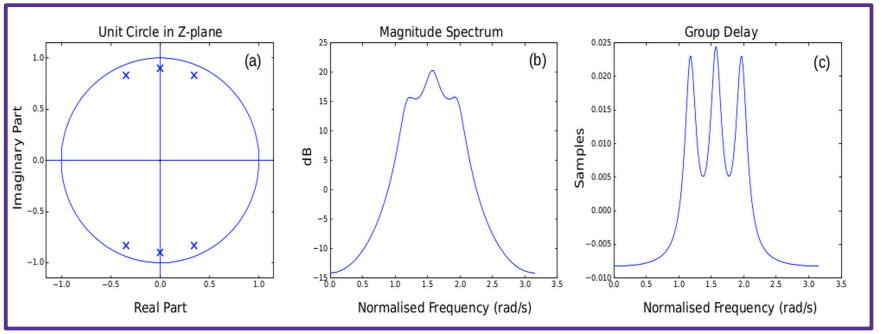


For more details please refer to ...

Loweimi, Erfan (2018) Robust Phase-based Speech Signal Processing From Source-Filter Separation to Model-Based Robust ASR. PhD thesis, University of Sheffield.

Group Delay (GD) (1)

- Negative spectral derivative of phase spectrum
- Advantages ...
  1) Additivity → $x(t) * y(t) \equiv GD_X(\omega) + GD_Y(\omega)$
  2) High spectral resolution



Group Delay (2)

- Advantages ...
  - 3) Similarity to mag spectrum for MinPhase signals
    - Similarity → max@poles & min@zeros
    - |X(ω)| is replaceable with GD in the pipeline + some amendments