SpeechWave

Speech Acoustic Modelling from Raw Phase Spectrum

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

[1] The 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

ABSTRACT

GOAL: Acoustic modeling using speech *raw phase* spectrum

* Raw: using entire spectrum (frequency ≥ 0)

How: using single-head and multi-head CNNs

CONTRIBUTIONS:

1. Acoustic modelling using raw phase spectrum
   – Bypass feature engineering
2. "Separate + Recombine" raw phase spectra of the source and filter components
3. Investigate fusion at different levels of abstraction
4. Study usefulness of the phase spectrum in a LVCSR task

Tasks: TIMIT (phone recognition) and WSJ (LVCSR)

RESULTS: Comparable to better WER than standard features

FUTURE WORK: Proposed framework is general; applicable to a wide range of speech recognition/classification tasks

Usefulness/Applications of Speech Phase Spectrum

Phase spectrum is not an appealing part of Fourier Transform

✗ Ambiguous shape, perceptual usefulness (?), ...

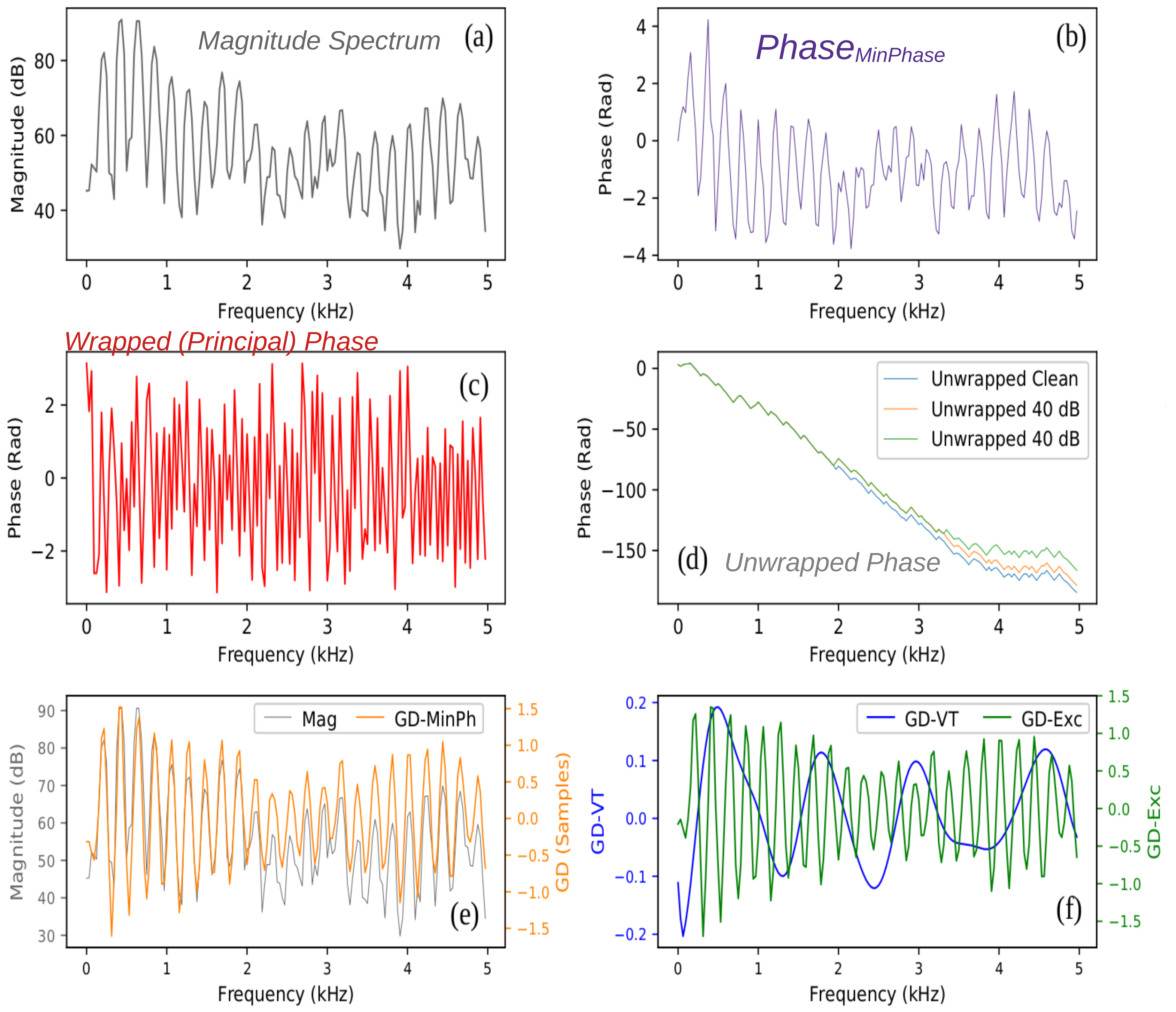
Applications:

✔ Speech analysis, enhancement, feature extraction, ...

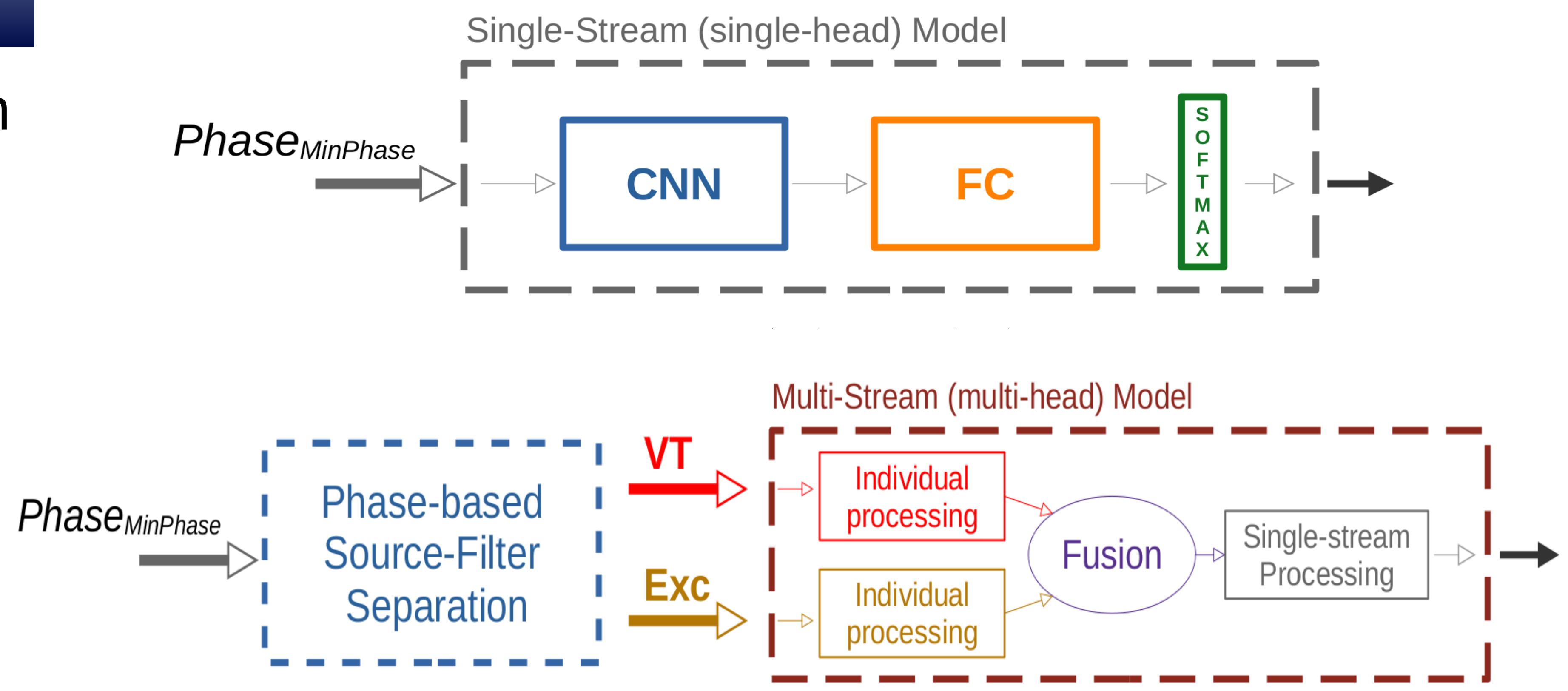
* For a detailed discussion and literature review please refer to ...

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

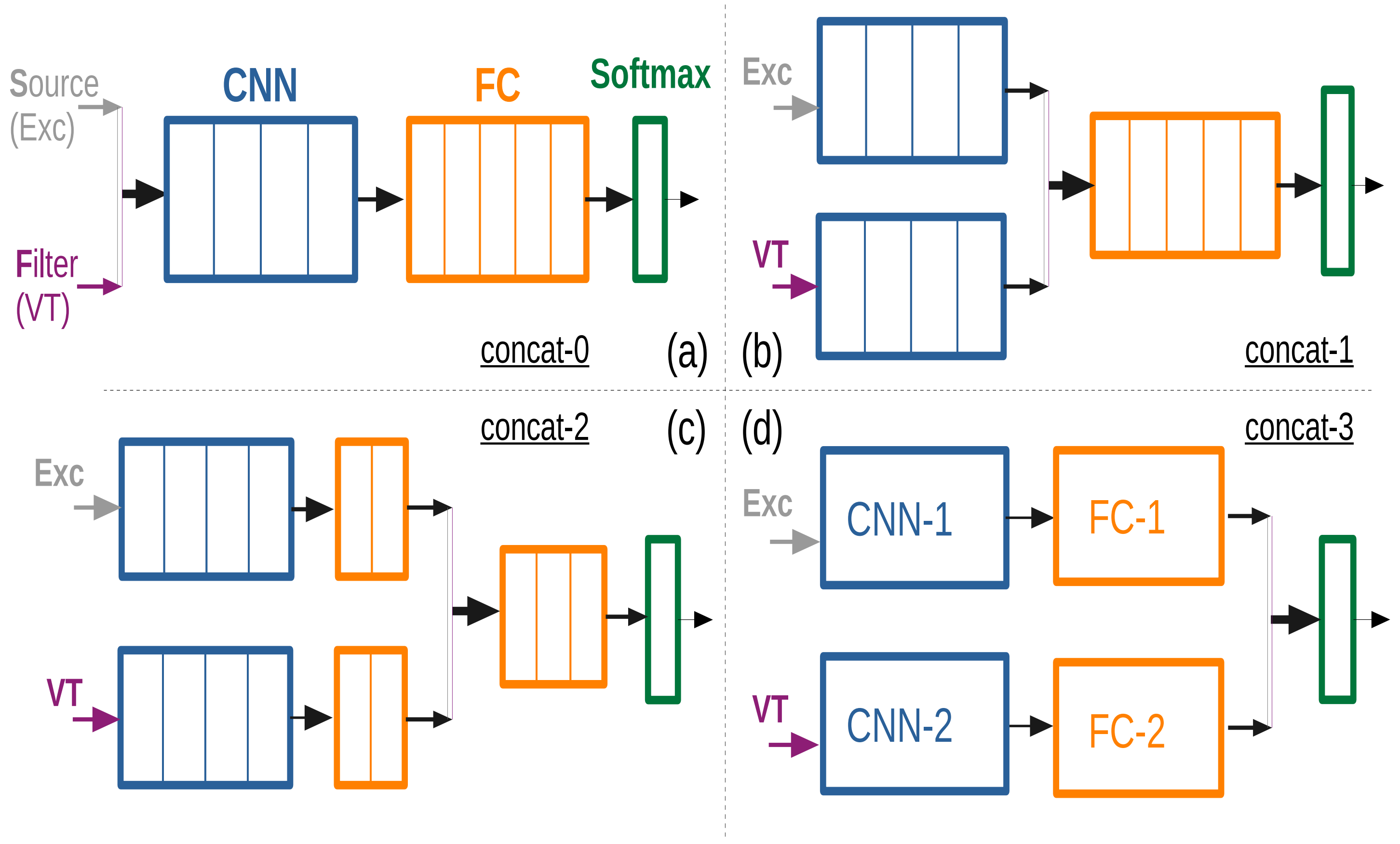
Raw Phase-based Representations



Architecture: Single-head vs Multi-head



Multi-stream Processing w/ Multiple Fusion Schemes



* **Advantages** ...

(1) Each info stream is weighted/gated properly

(2) Bespoke transforms for each info stream learned

(3) Fusion at optimal abstraction level (potentially)

Experimental Results

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$^{0.1}$	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

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
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

(1) Phase-based features outperform mag-based ones

(2) Even for Wrapped phase, decent results achieved

(3) Multi-stream (multi-head) outperforms single-stream

(4) Optimal Fusion level is Concat-1