Source-filter Separation of Speech Signal in the Phase Domain

Erfan Loweimi
Jon Barker
Thomas Hain

July, 2015

Outline

- Problems with phase spectrum
- Group delay function (GDF)
- Phase information content
- Speech signal decomposition
- Phase-based source-filter separation
- Feature extraction for ASR
- Conclusion

Problems

Challenges

- Historical Considerations

Challenges

- Historical Considerations
  - Ohm's acoustic law (1843) + Helmholtz (1875)

Challenges

- Historical Considerations
  - Ohm's acoustic law (1843) + Helmholtz (1875)
    - “the percepted quality of a tone depends solely on the *number* and *relative strength* of its partial simple tones, and not on their relative phases”

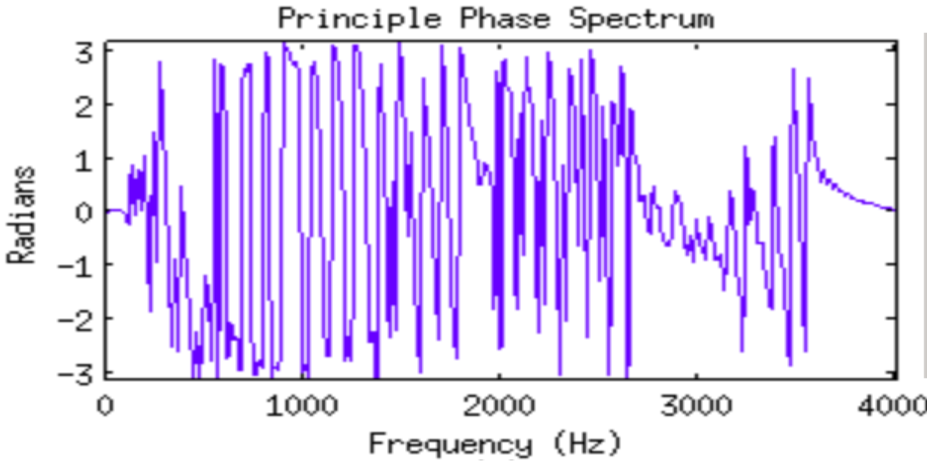
Challenges

- Historical Considerations
  - Ohm's acoustic law (1843) + Helmholtz (1875)
    - “the percepted quality of a tone depends solely on the *number* and *relative strength* of its partial simple tones, and not on their relative phases”

  - Although some studies show that the auditory system is not totally “*phase deaf*”, this law forms the status qua

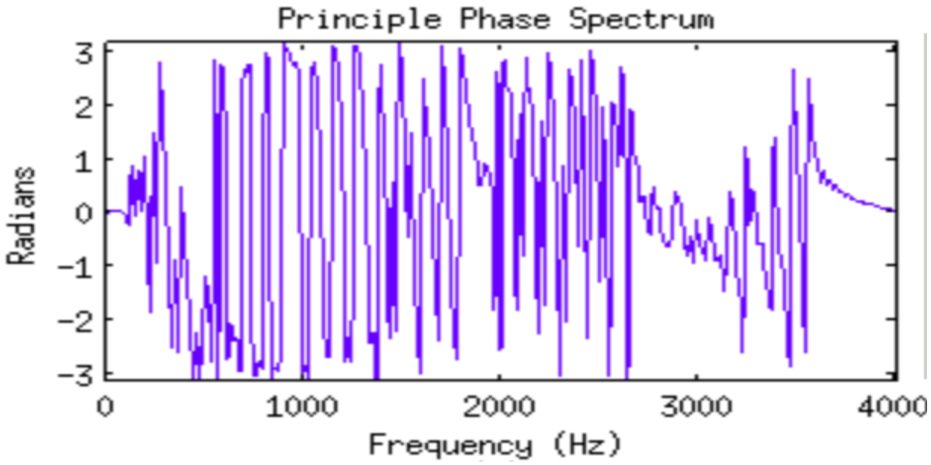
Challenges ...

- Phase wrapping



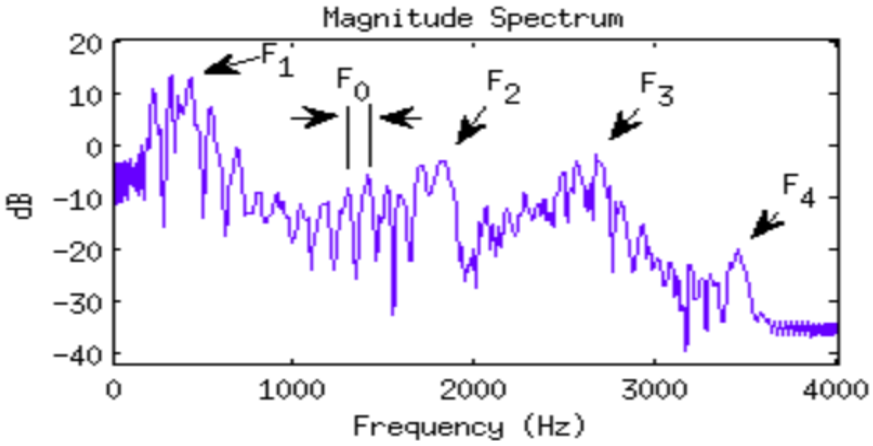
Challenges ...

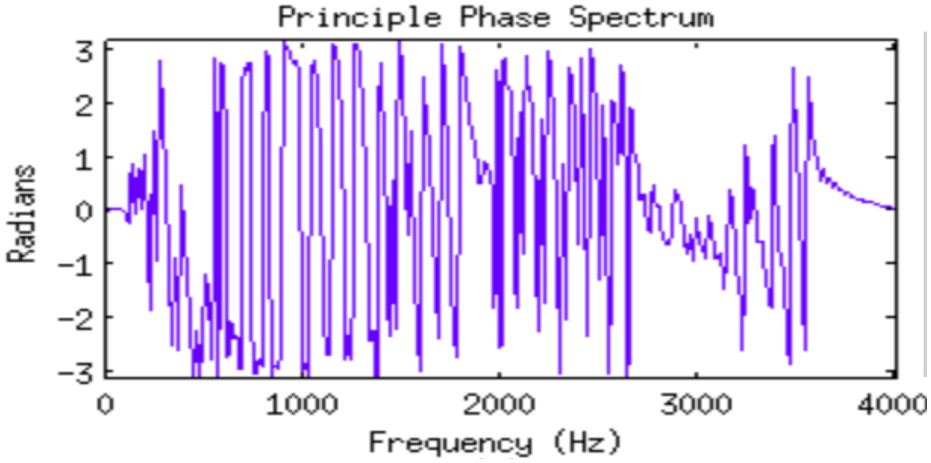
- Phase wrapping
  - Chaotic/noise-like behaviour
  - Lacks any meaningful trend or extrema points
    - ✖ Physical interpretation
    - ✖ Mathematical modelling



Challenges ...

- Phase wrapping
  - Chaotic/noise-like behaviour
  - Lacks any meaningful trend or extrema points
    - Physical interpretation
    - Mathematical modelling





Challenges ...

- Only informative in long-term (> x00 ms)
  - Violates stationarity assumption !
  - In short frames (~ 30 ms), it is generally believed that the phase spectrum does not contribute much to speech quality/intelligibility

Group Delay Function (GDF)

- Definition

$$\tau_X(\omega) = -\frac{d}{d\omega} arg[X(\omega)] = -Im\{\frac{d}{d\omega}\log(X(\omega))\}$$

Group Delay Function (GDF)

- Definition

$$\tau_X(\omega) = -\frac{d}{d\omega} arg[X(\omega)] = -Im\{\frac{d}{d\omega}\log(X(\omega))\}$$

$$\tau_X(\omega) = \frac{X_{Re}(\omega)Y_{Re}(\omega) + X_{Im}(\omega)Y_{Im}(\omega)}{|X(\omega)|^2}$$

Group Delay Function (GDF)

- Definition

$$\tau_X(\omega) = -\frac{d}{d\omega} arg[X(\omega)] = -Im\{\frac{d}{d\omega}\log(X(\omega))\}$$

$$\tau_X(\omega) = \frac{X_{Re}(\omega)Y_{Re}(\omega) + X_{Im}(\omega)Y_{Im}(\omega)}{|X(\omega)|^2}$$

- Pros
  - Resembles the magnitude spectrum
  - High frequency resolution
  - Additive

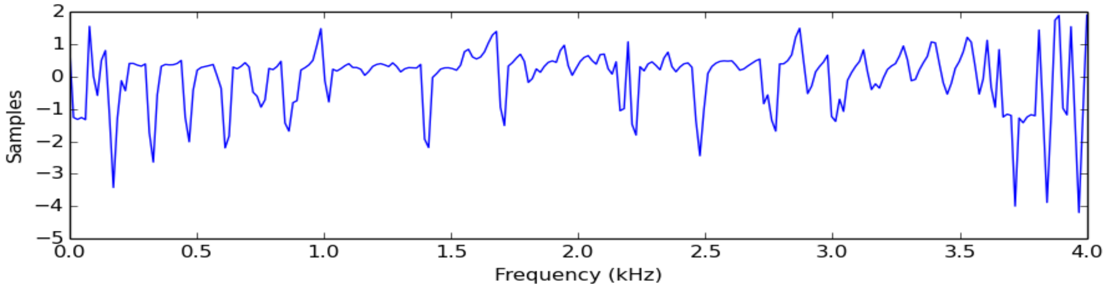
Group Delay Function (GDF)

- Definition

$$\tau_X(\omega) = -\frac{d}{d\omega} arg[X(\omega)] = -Im\{\frac{d}{d\omega}\log(X(\omega))\}$$

$$\tau_X(\omega) = \frac{X_{Re}(\omega)Y_{Re}(\omega) + X_{Im}(\omega)Y_{Im}(\omega)}{|X(\omega)|^2}$$

- Pros
  - Resembles the magnitude spectrum
  - High frequency resolution
  - Additive
- Cons
  - Too spiky



Phase Information Content

- What is the information?

Phase Information Content

- What is the information?
  - Context dependent
    - Information theory: average of uncertainty
    - Speech: lingual content, speaker ID, ...

Phase Information Content

- What is the information?
  - Context dependent
    - Information theory: average of uncertainty
    - Speech: lingual content, speaker ID, ...
- Is phase informative?
  - ✓ From perceptual viewpoint
  - ✗ From signal processing viewpoint

Signal Decomposition

- For any signal

$$X(\omega) = |X(\omega)| \,.\, e^{j\phi_X(\omega)}$$

$$\begin{aligned} X(\omega) &= X_{MinPh}(\omega) \,.\, X_{AllPass}(\omega) \\ &= |X_{MinPh}(\omega)| e^{j\phi_{MinPh}(\omega)} \,.\, 1 e^{j\phi_{AllPass}(\omega)} \end{aligned}$$

Signal Decomposition

- For any signal

$$X(\omega) = |X(\omega)| \,.\, e^{j\phi_X(\omega)}$$

$$\begin{aligned} X(\omega) &= X_{MinPh}(\omega) \,.\, X_{AllPass}(\omega) \\ &= |X_{MinPh}(\omega)| e^{(j\phi_{MinPh}(\omega)+\phi_{AllPass}(\omega))} \end{aligned}$$

$$\begin{cases} |X(\omega)| = |X_{MinPh}(\omega)| \\ \phi_X(\omega) = \phi_{MinPh}(\omega) + \phi_{AllPass}(\omega) \end{cases}$$

Signal Decomposition

- For any signal

$$\begin{cases} |X(\omega)| = |X_{MinPh}(\omega)| \\ \phi_X(\omega) = \phi_{MinPh}(\omega) + \phi_{AllPass}(\omega) \end{cases}$$

* Is there any relation between phase and magnitude spectra?

Signal Decomposition

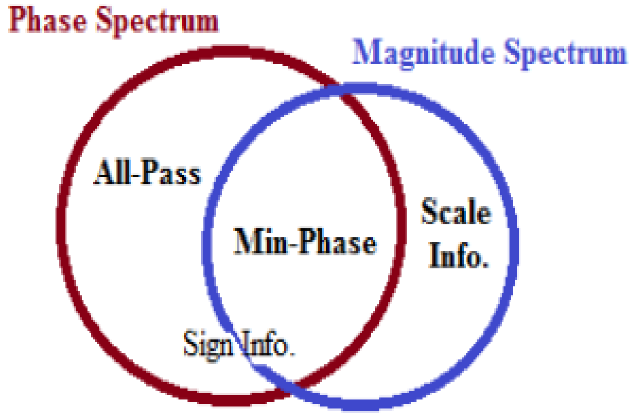
- For any signal

$$\begin{cases} |X(\omega)| = |X_{MinPh}(\omega)| \\ \phi_X(\omega) = \phi_{MinPh}(\omega) + \phi_{AllPass}(\omega) \end{cases}$$

* Is there any relation between phase and magnitude spectra?

$$|X_{MinPh}(\omega)| \quad \underleftrightarrow{\text{HiL.Tran}} \quad \phi_{MinPh}(\omega)$$

Signal Decomposition



- For any signal

$$\begin{cases} |X(\omega)| = |X_{MinPh}(\omega)| \\ \phi_X(\omega) = \phi_{MinPh}(\omega) + \phi_{AllPass}(\omega) \end{cases}$$

* Is there any relation between phase and magnitude spectra?

$$|X_{MinPh}(\omega)| \quad \underleftrightarrow{\text{HiL.Tran}} \quad \phi_{MinPh}(\omega)$$

Signal Decomposition ...

- For speech ...

$$\begin{cases} |X(\omega)| = |X_{VT}(\omega)|.|X_{Exc}(\omega)| \\ \\ \end{cases}$$

Signal Decomposition ...

- For speech ...

$$\begin{cases} |X(\omega)| = |X_{VT}(\omega)|.|X_{Exc}(\omega)| = |X_{MinPh}(\omega)| \\ \end{cases}$$

Signal Decomposition ...

- For speech

$$\begin{cases} |X(\omega)| = |X_{VT}(\omega)|.|X_{Exc}(\omega)| = |X_{MinPh}(\omega)| \\ |X_{MinPh}(\omega)| \quad \xleftrightarrow{\text{HiL.Tran}} \quad \arg[X_{MinPh}(\omega)] \end{cases}$$

Signal Decomposition ...

- For speech

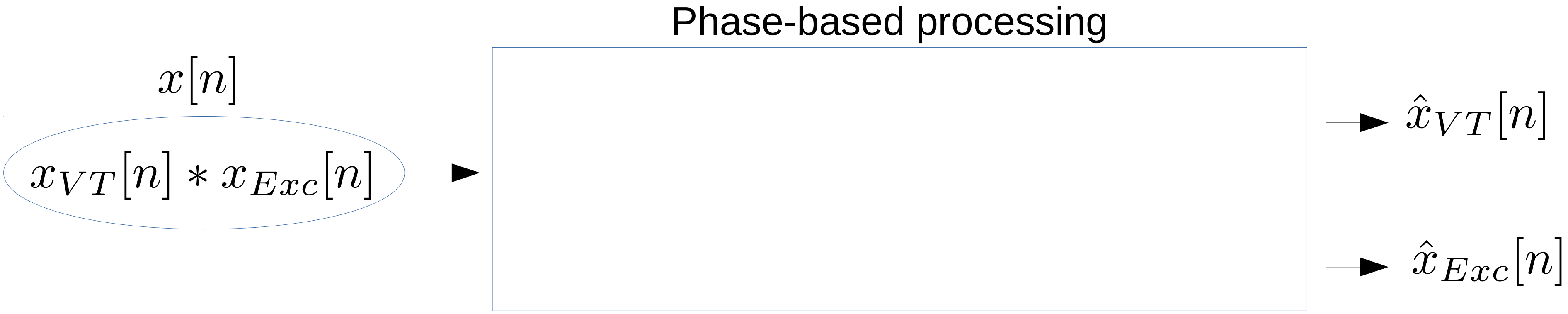
$$\begin{cases} |X(\omega)| = |X_{VT}(\omega)|.|X_{Exc}(\omega)| = |X_{MinPh}(\omega)| \\ |X_{MinPh}(\omega)| \quad \underleftrightarrow{\text{HiL.Tran}} \quad \arg[X_{MinPh}(\omega)] \\ arg[X_{MinPh}(\omega)] = arg[X_{VT}(\omega)] + arg[X_{Exc}(\omega)]. \end{cases}$$

Signal Decomposition ...

- For speech

$$\begin{cases} |X(\omega)| = |X_{VT}(\omega)|.|X_{Exc}(\omega)| = |X_{MinPh}(\omega)| \\ |X_{MinPh}(\omega)| \quad \xleftrightarrow{\text{HiL.Tran}} \quad \arg[X_{MinPh}(\omega)] \\ arg[X_{MinPh}(\omega)] = arg[X_{VT}(\omega)] + arg[X_{Exc}(\omega)]. \end{cases}$$

- Goal ...

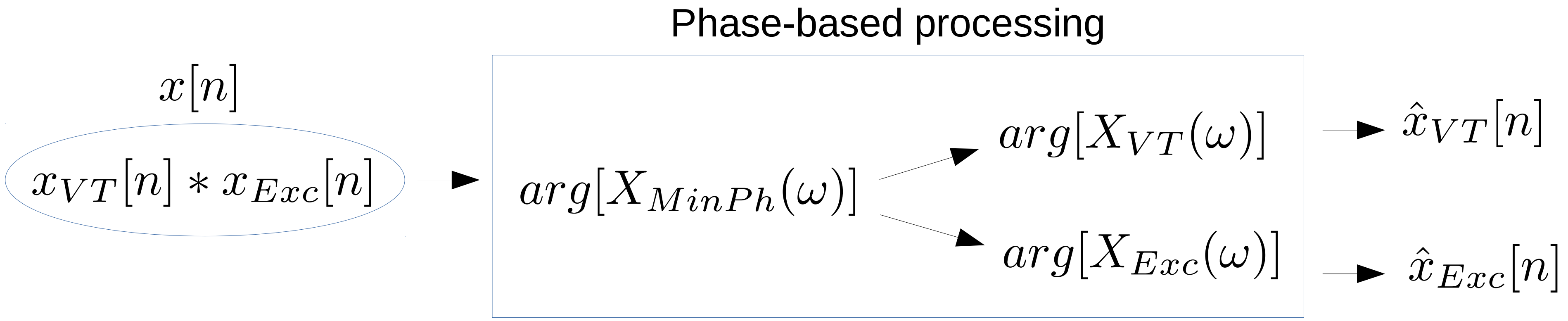


Signal Decomposition ...

- For speech

$$\begin{cases} |X(\omega)| = |X_{VT}(\omega)|.|X_{Exc}(\omega)| = |X_{MinPh}(\omega)| \\ |X_{MinPh}(\omega)| \quad \underset{\longleftrightarrow}{\text{HiL.Tran}} \quad \arg[X_{MinPh}(\omega)] \\ arg[X_{MinPh}(\omega)] = arg[X_{VT}(\omega)] + arg[X_{Exc}(\omega)]. \end{cases}$$

- Goal ...



Phase-based processing

$x[n]$

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

$arg[X_{MinPh}(\omega)]$

$arg[X_{VT}(\omega)]$

$arg[X_{Exc}(\omega)]$

$\hat{x}_{VT}[n]$

$\hat{x}_{Exc}[n]$

MinPh Component Computation

- In Frequency domain

- In Quefrency domain

MinPh Component Computation

- In Frequency domain

$$\begin{aligned}\arg[\mathrm{X}_{MinPh}(\omega)] &= Hil\{log|X_{MinPh}(\omega)|\} \\ &= -\frac{1}{2\pi}log|X_{MinPh}(\omega)| * cot(\frac{\omega}{2})\end{aligned}$$

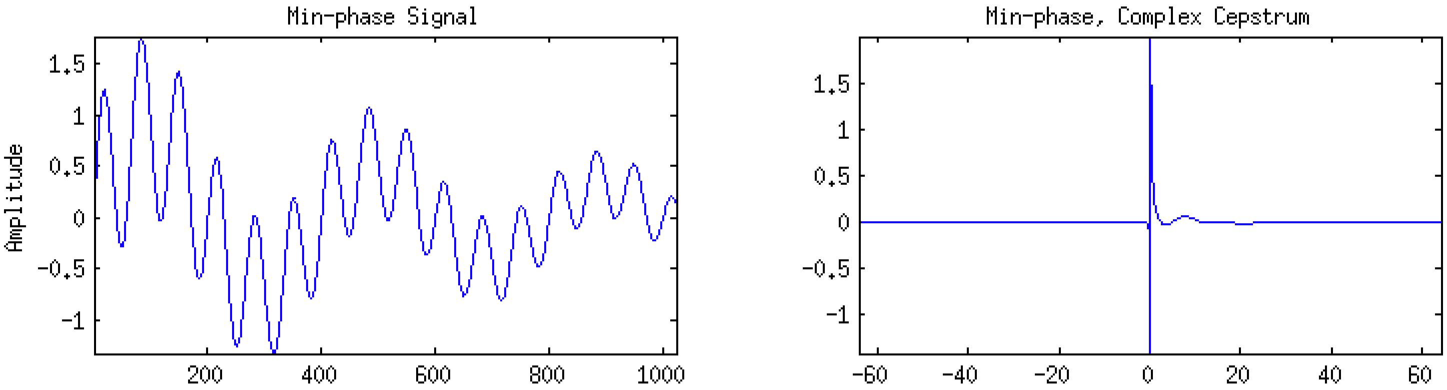
- In Quefrency domain
  - Apply a proper lifter on the complex cepstrum

MinPh Component Computation

- In Frequency domain

$$\begin{aligned}\arg[\mathrm{X}_{MinPh}(\omega)] &= Hil\{log|X_{MinPh}(\omega)|\}\\ &= -\frac{1}{2\pi}log|X_{MinPh}(\omega)| * cot(\frac{\omega}{2})\end{aligned}$$

- In Quefrency domain
  - Apply a proper lifter on the complex cepstrum



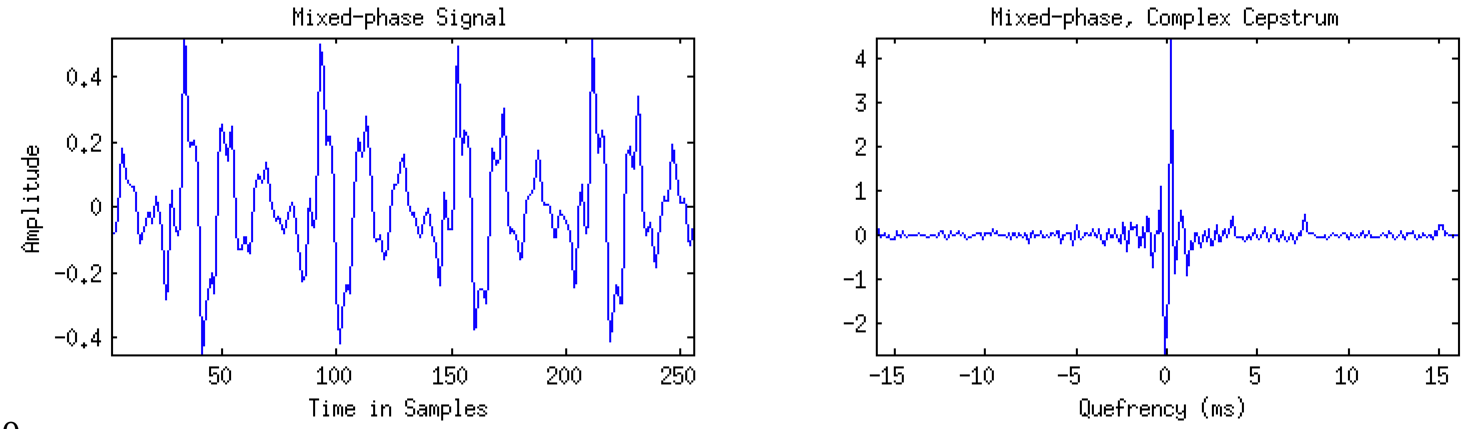
MinPh sequence => causal cepstrum

MinPh Component Computation

- In Frequency domain

$$\begin{aligned}\arg[\mathrm{X}_{MinPh}(\omega)] &= Hil\{log|X_{MinPh}(\omega)|\} \\ &= -\frac{1}{2\pi} log|X_{MinPh}(\omega)| * cot(\frac{\omega}{2})\end{aligned}$$

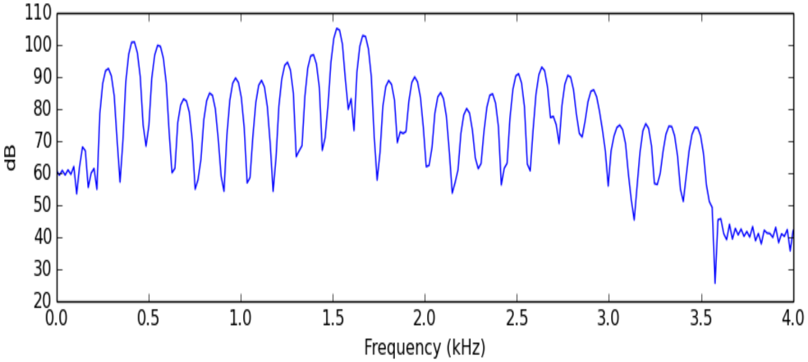
- In Quefrency domain
  - Apply a proper lifter on the complex cepstrum



Speech is mixed-phase

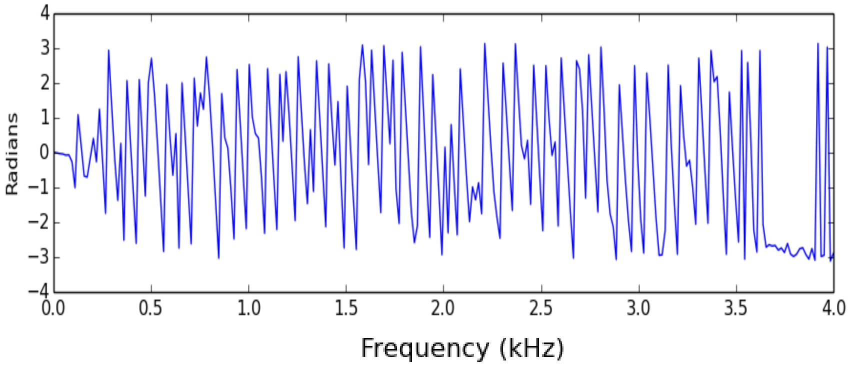
MinPhase component

$log|X(\omega)|$



Hilbert Transform

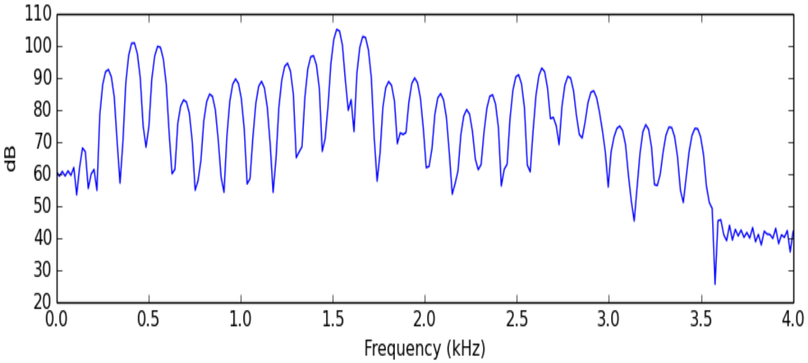
$ARG[X(\omega)]$



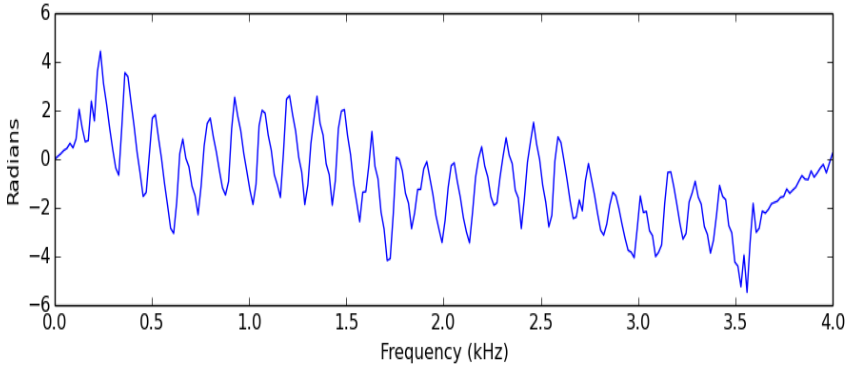
$arg[X_{MinPh}(\omega)]$

MinPhase component

$log|X(\omega)|$

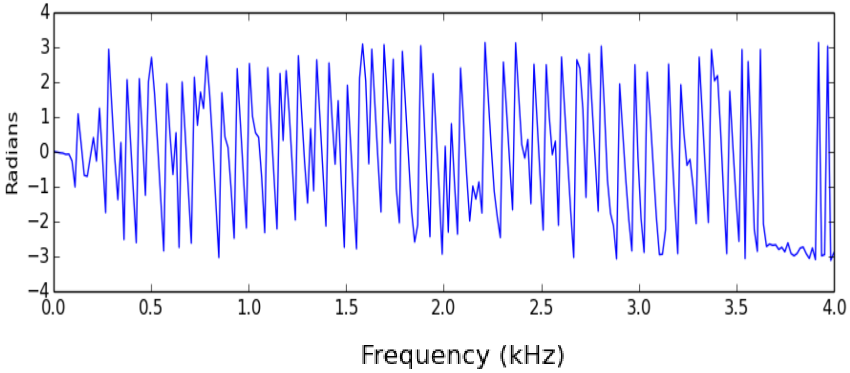


Hilbert Transform

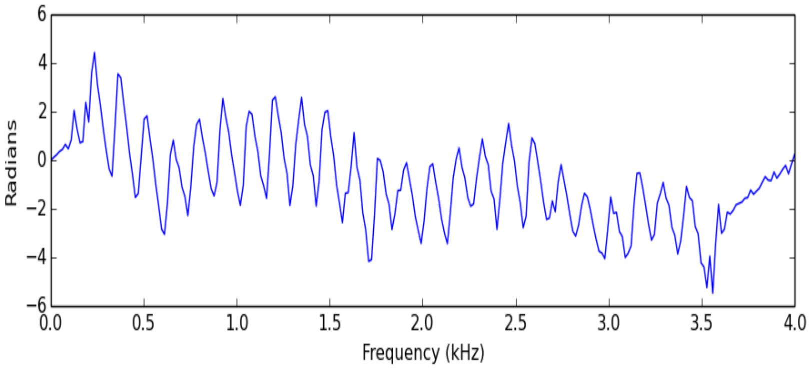


$arg[X_{MinPh}(\omega)]$

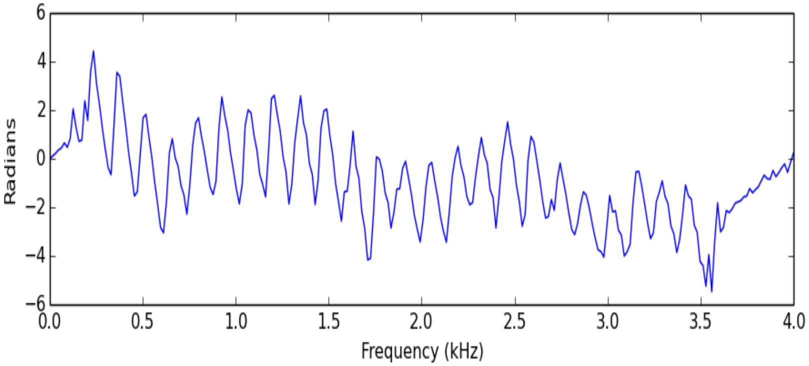
$ARG[X(\omega)]$



Trend/Fluctuation Analysis

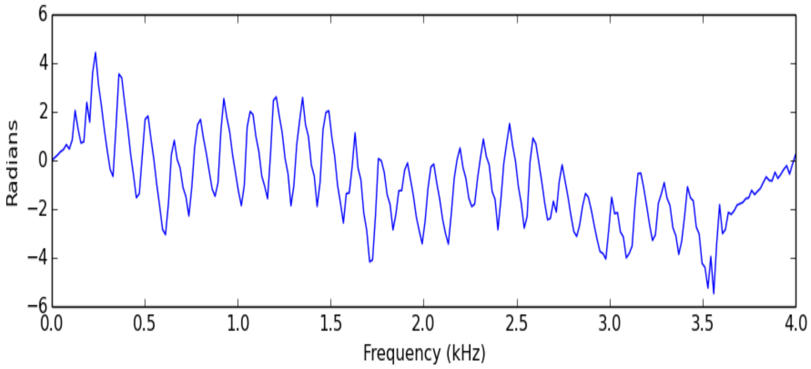


Trend/Fluctuation Analysis

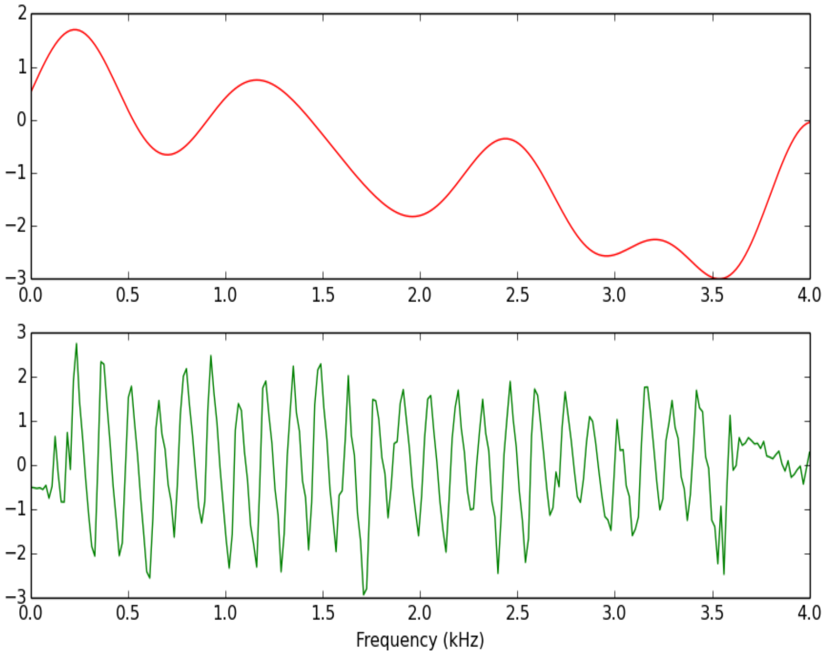


$$\longrightarrow \; arg[X_{MinPh}] = \; Trend \; + \; Fluctuation$$

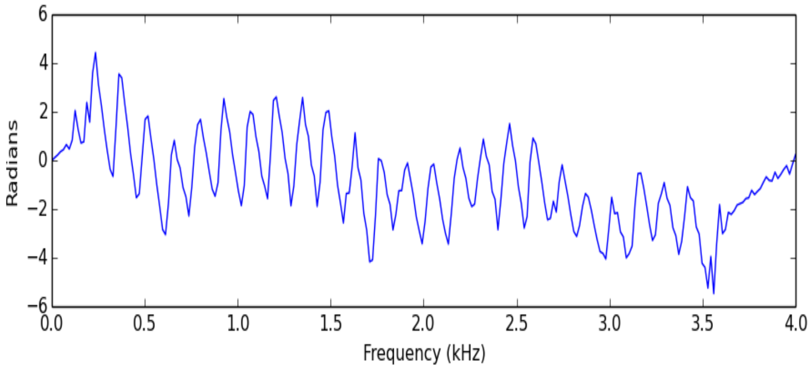
Trend/Fluctuation Analysis



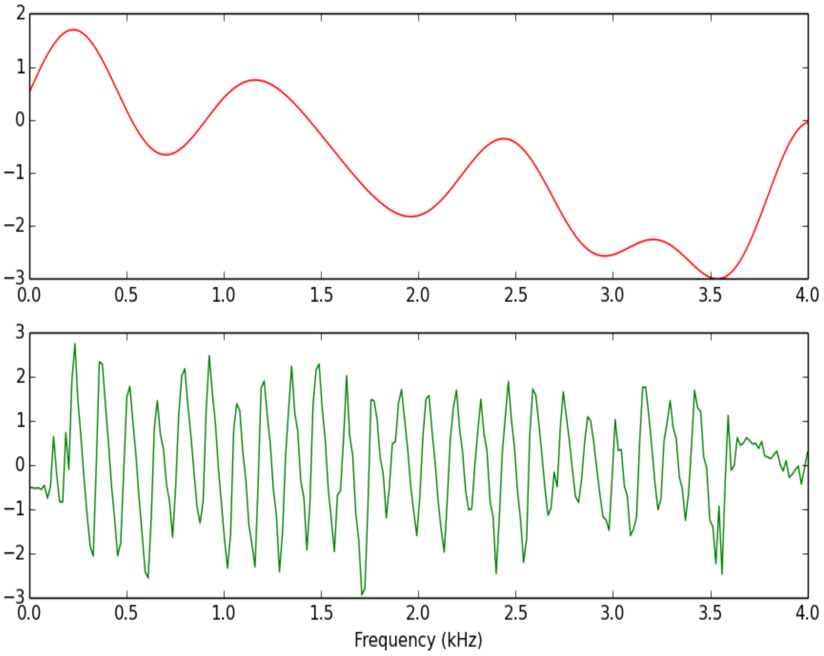
$$arg[X_{MinPh}] = Trend + Fluctuation$$



Trend/Fluctuation Analysis



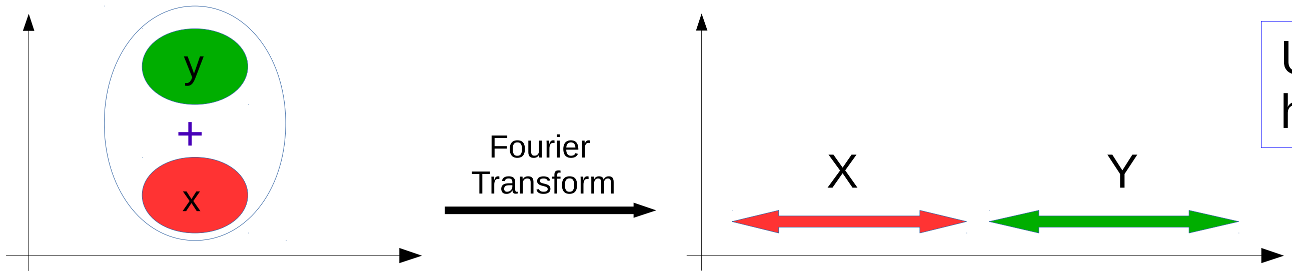
$$arg[X_{MinPh}] = \color{red}{Trend} + \color{green}{Fluctuation}$$



Vocal Tract (Filter)

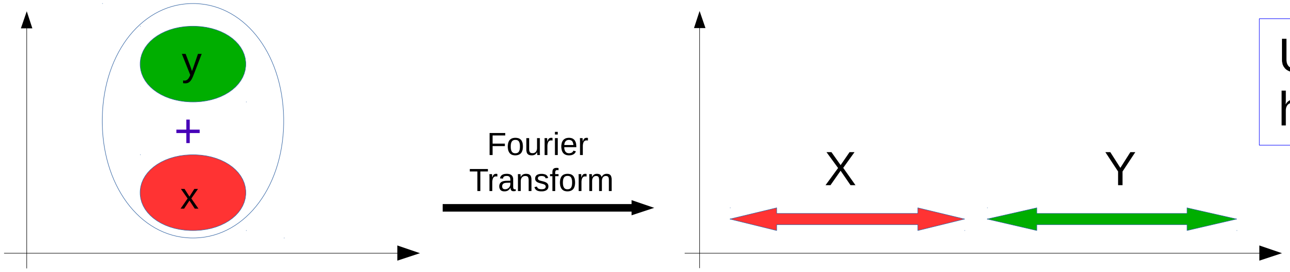
Excitation (Source)

Trend/Fluctuation Separation

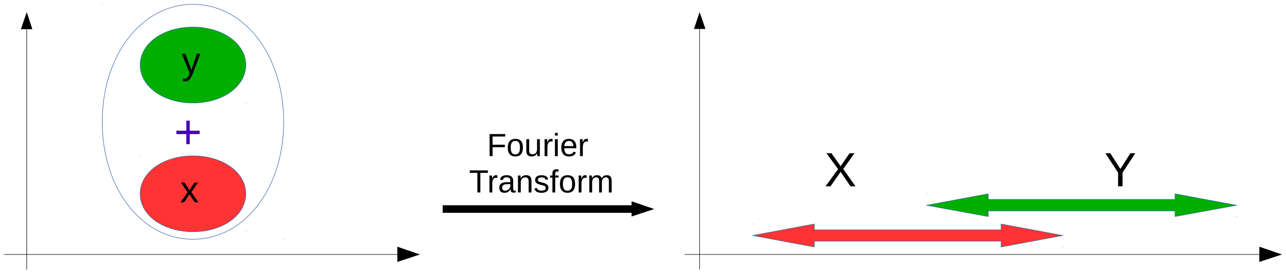


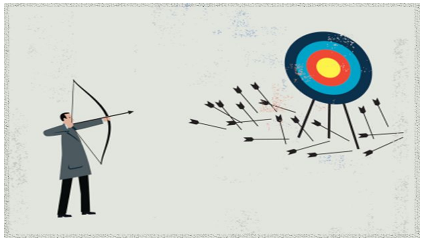
Underlying assumption for having successful separation

Trend/Fluctuation Separation

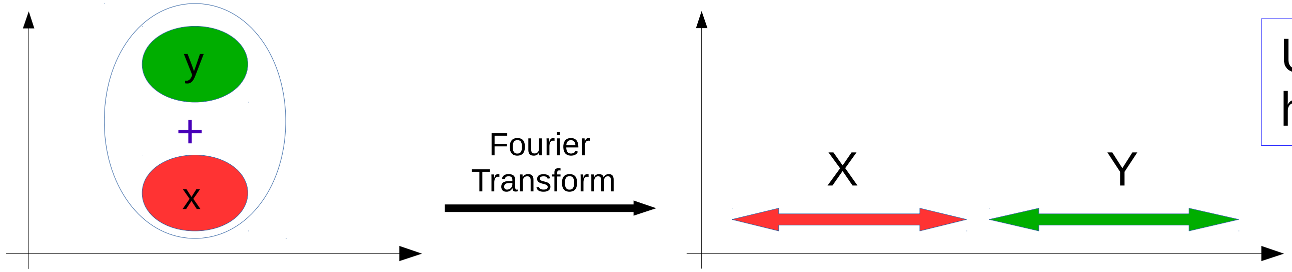


Underlying assumption for having successful separation

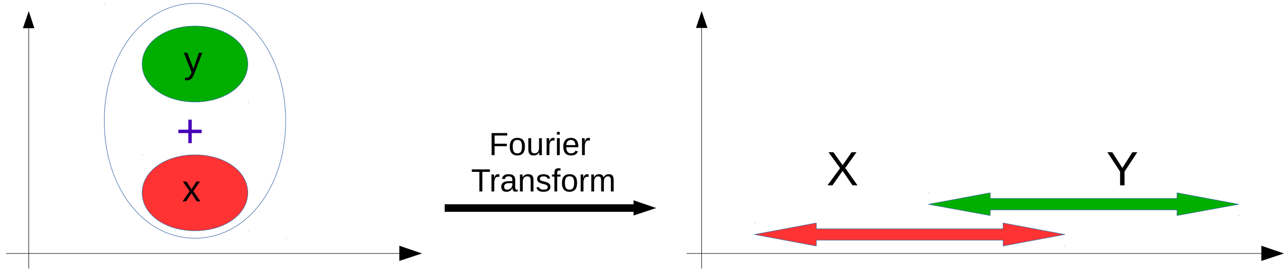


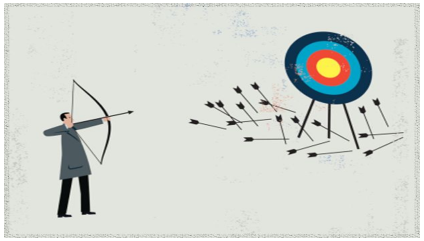


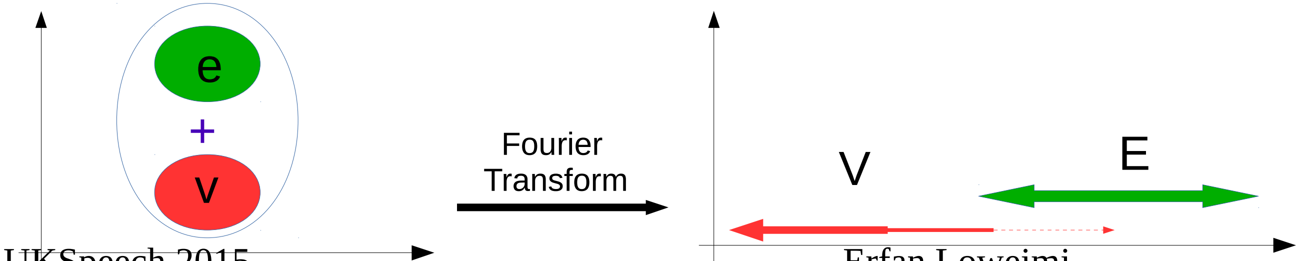
Trend/Fluctuation Separation



Underlying assumption for having successful separation

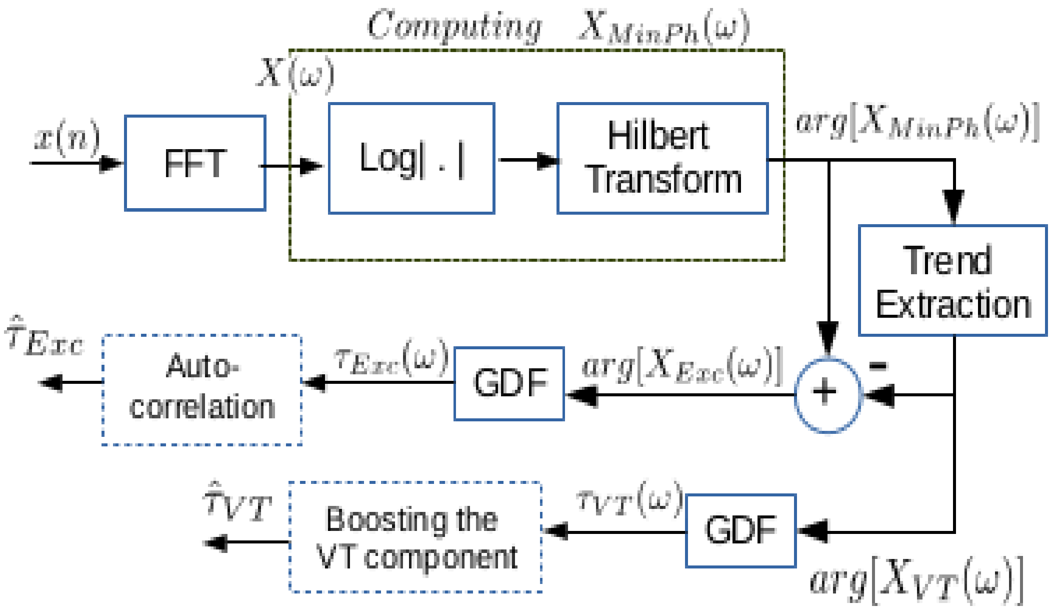




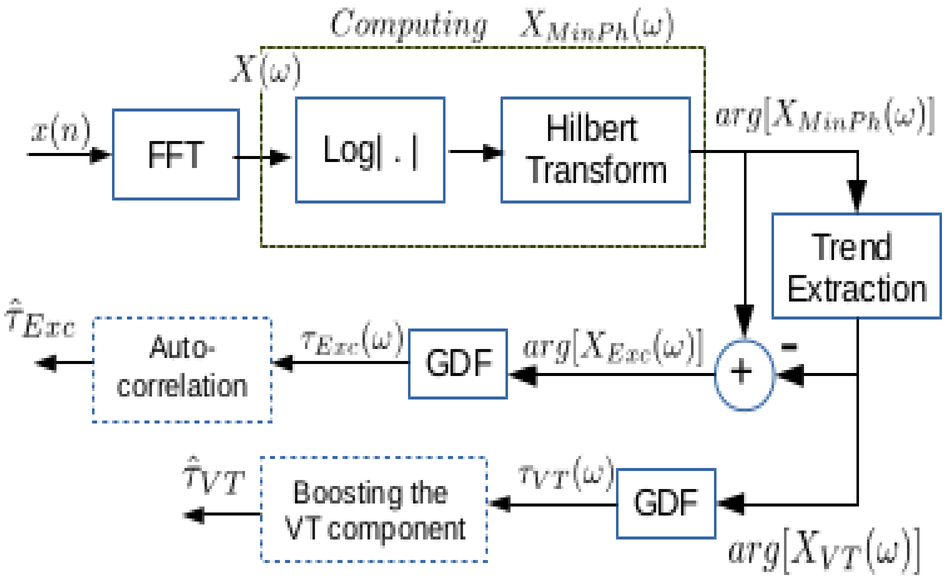


Speech signal

Phase-based Source-Filter Decomposition

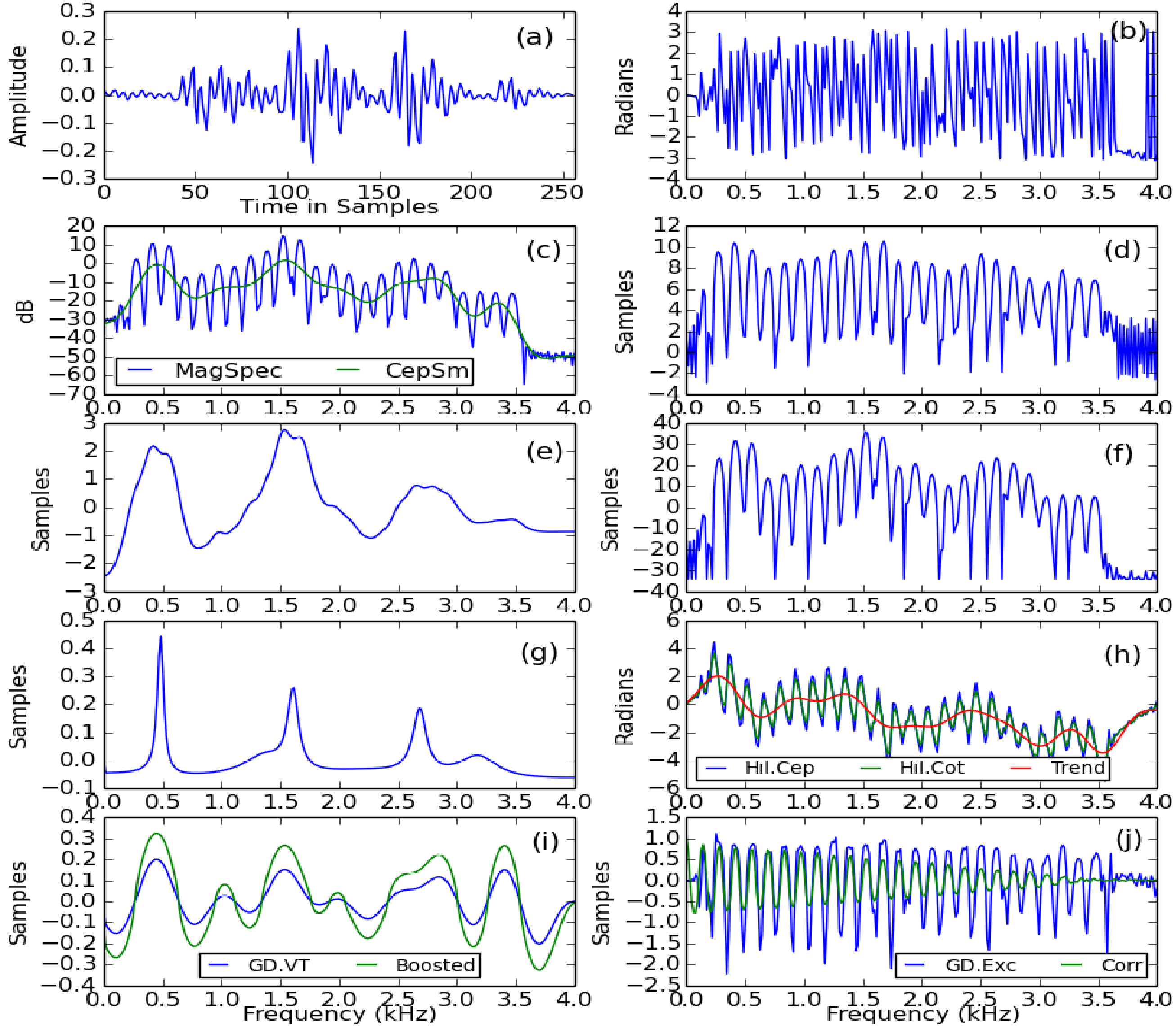


Phase-based Source-Filter Decomposition



$$\begin{cases} \hat{\tau}_{VT}(\omega) = signum(\tau_{VT}(\omega)).|\tau_{VT}(\omega)|^{\alpha} \\ signum(\tau_{VT}(\omega)) = \frac{\tau_{VT}(\omega)}{|\tau_{VT}(\omega)|} \end{cases}$$

Phase-based Source-Filter Decomposition



waveform

Phase Spec.

Mag. Spec.

MODGDF(03)

CGDF

Product Spec.

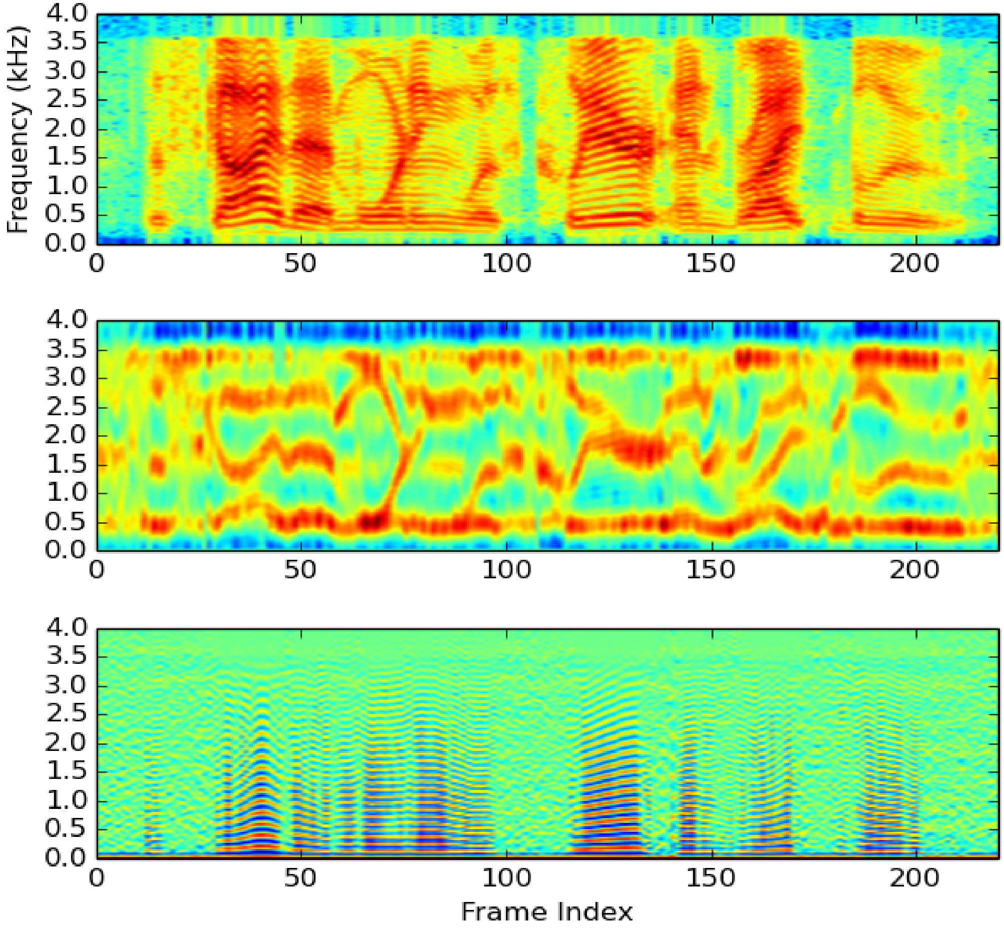
AR+GDF

MinPh Phase

GDF.VT

GDF.Exc

Phase-based Source-Filter Decomposition

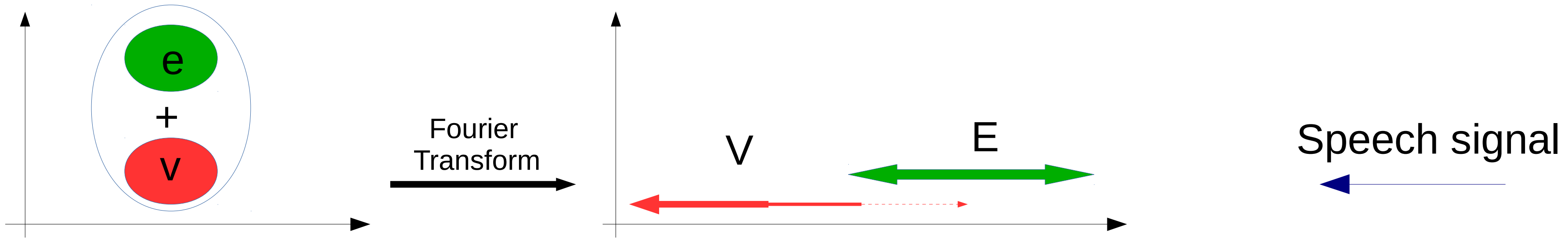


$log|X(n, \omega)|$

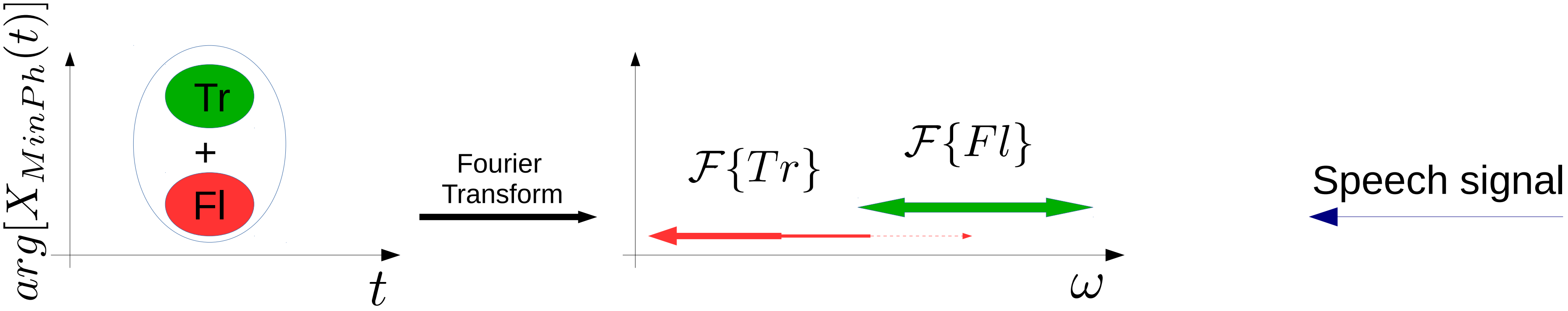
$\tau_{VT}(n, \omega)$

$\tau_{Exc}(n, \omega)$

Decomposition in GDF domain



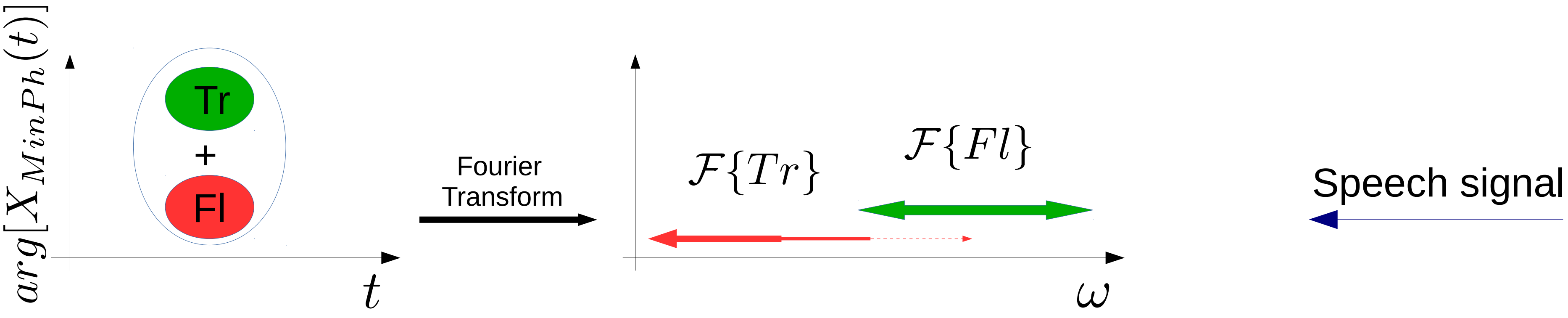
Decomposition in GDF domain



Fourier Transform

Speech signal

Decomposition in GDF domain



$arg[X_{MinPh}(t)]$

Tr

+

Fl

t

Fourier Transform

$\mathcal{F}\{Tr\}$

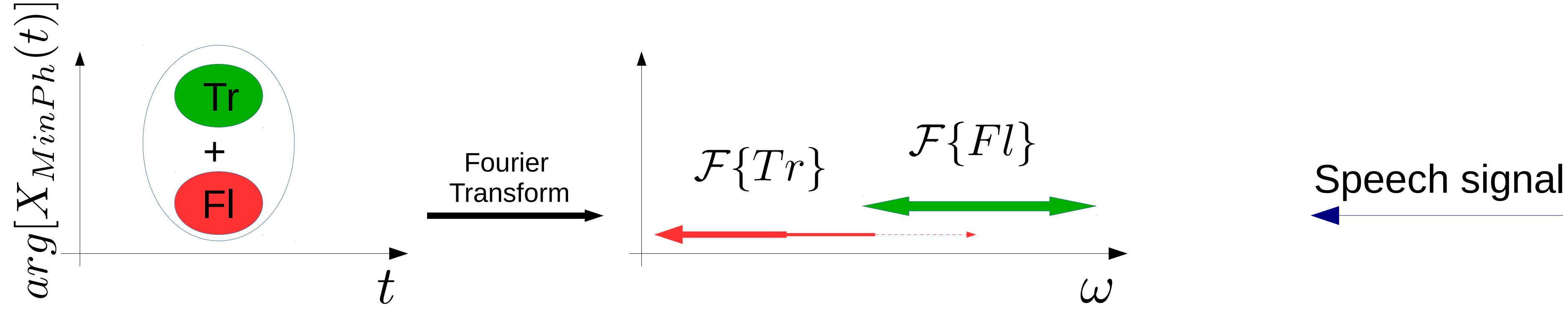
$\mathcal{F}\{Fl\}$

ω

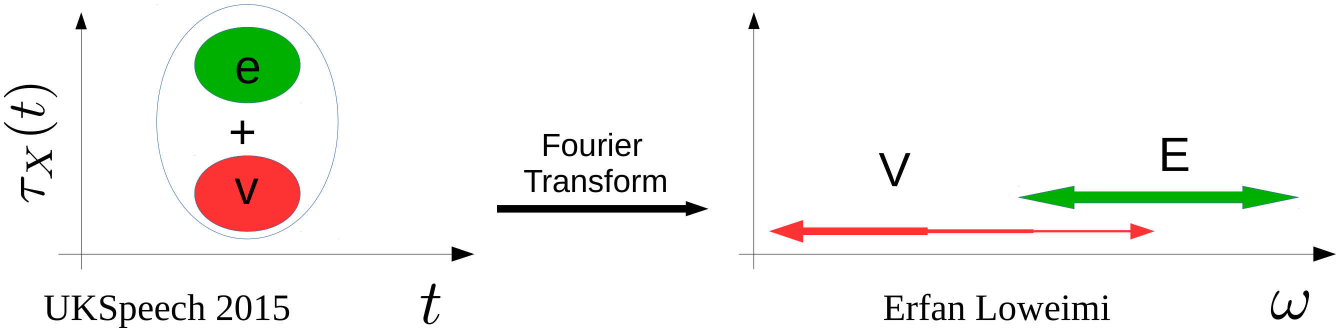
Speech signal

$$\begin{cases} \tau_X(t) = -\frac{d}{dt} arg[X_{MinPh}(t)] = -\frac{d}{dt} Trend - \frac{d}{dt} Fluctuation \\ \mathcal{F}\{\tau_X(t)\} = -j\omega \mathcal{F}\{Trend\} - j\omega \mathcal{F}\{Fluctuation\} \end{cases}$$

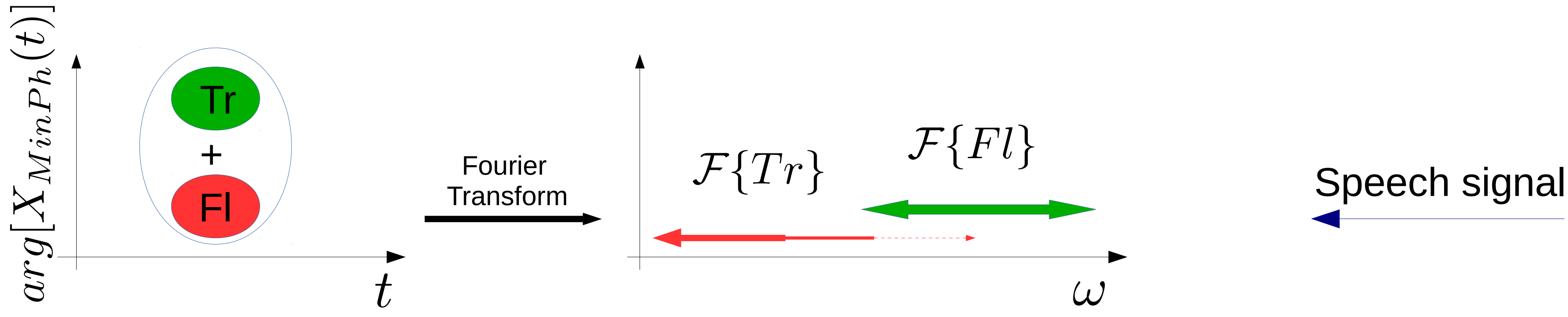
Decomposition in GDF domain



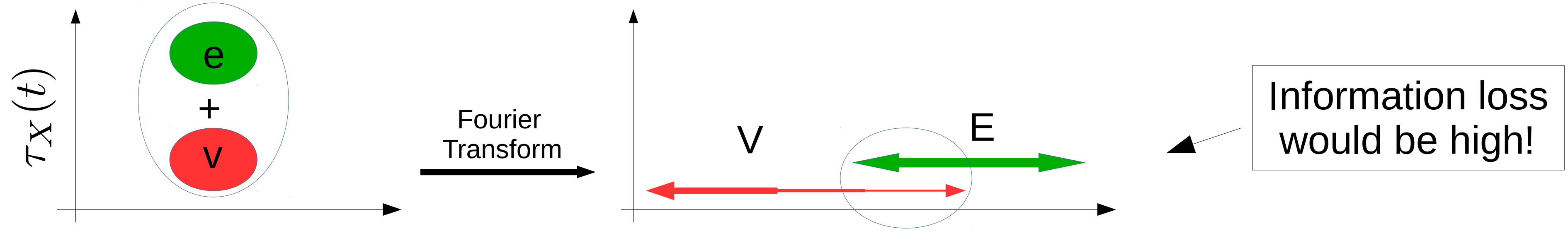
$$\begin{cases} \tau_X(t) = -\frac{d}{dt}arg[X_{MinPh}(t)] = -\frac{d}{dt}Trend - \frac{d}{dt}Fluctuation \\ \mathcal{F}\{\tau_X(t)\} = -j\omega\mathcal{F}\{Trend\} - j\omega\mathcal{F}\{Fluctuation\} \end{cases}$$



Decomposition in GDF domain



$$\begin{cases} \tau_X(t) = -\frac{d}{dt}arg[X_{MinPh}(t)] = -\frac{d}{dt}Trend - \frac{d}{dt}Fluctuation \\ \mathcal{F}\{\tau_X(t)\} = -j\omega\mathcal{F}\{Trend\} - j\omega\mathcal{F}\{Fluctuation\} \end{cases}$$



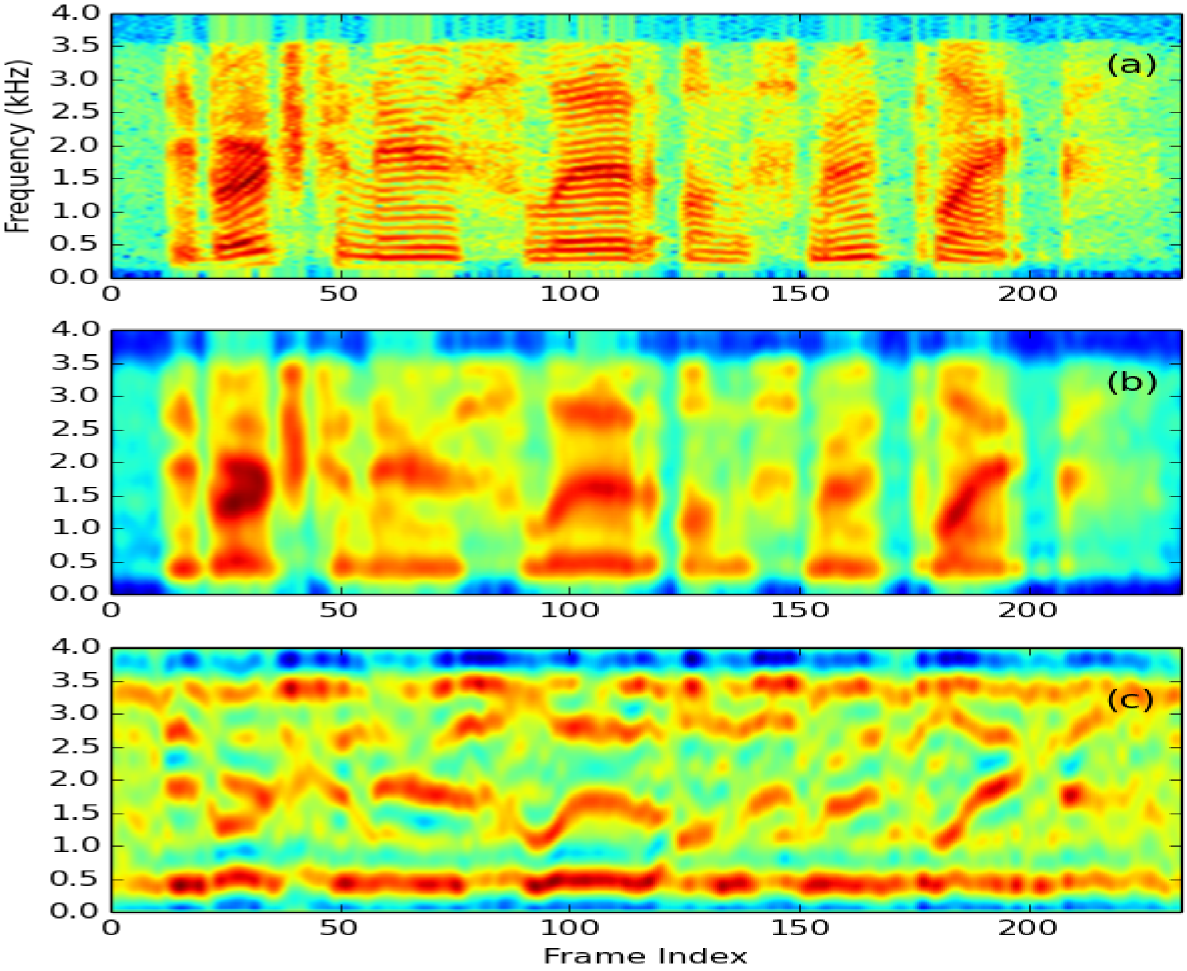
Information loss would be high!

Decomposition in log-magnitude domain

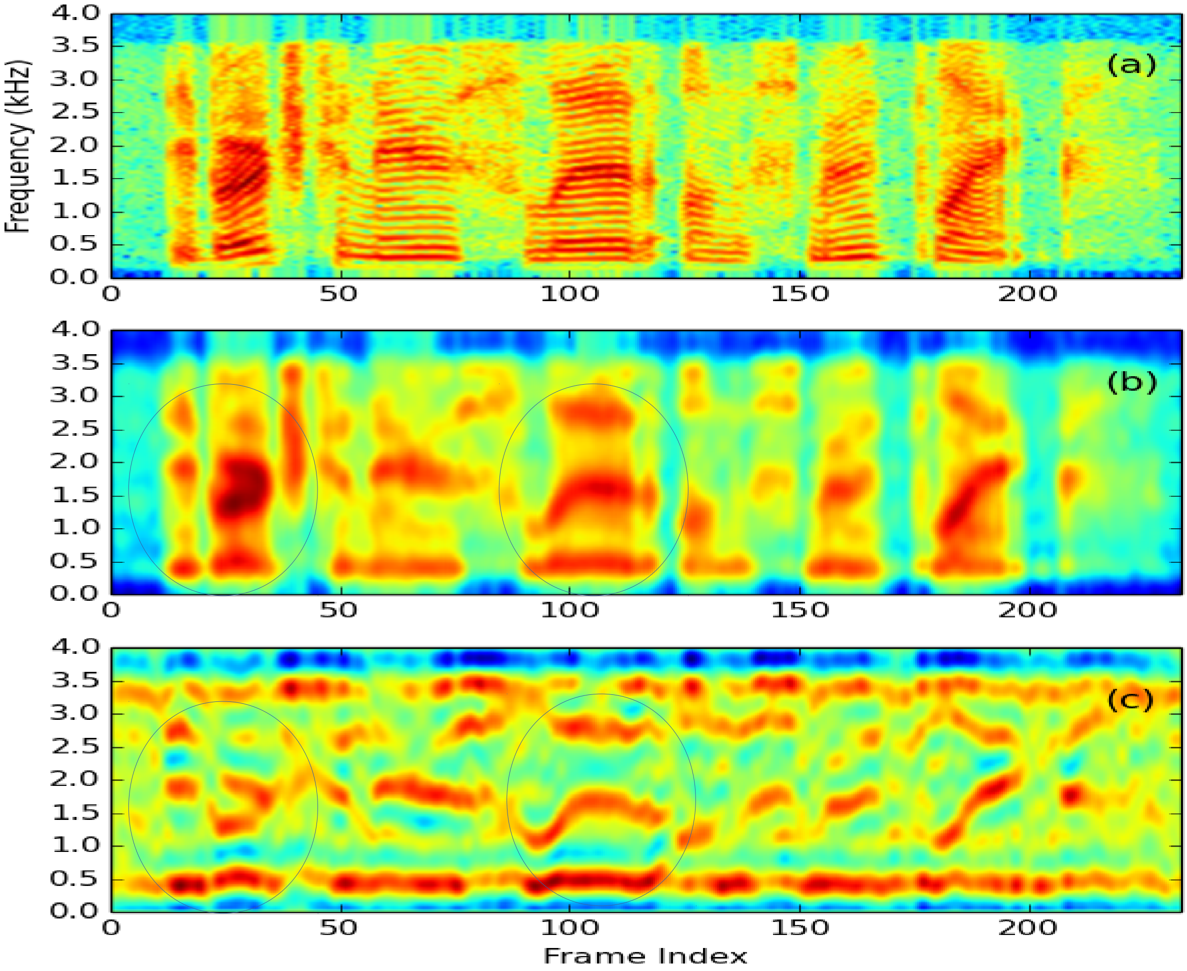
$log|X(n,\omega)|$

Smoothed Spec.

$\tau_X(n,\omega)$



Decomposition in log-magnitude domain



$log|X(n,\omega)|$

Smoothed Spec.

$\tau_X(n,\omega)$

Feature Extraction for ASR

i) $arg[X_{VT}] \rightarrow DCT \Rightarrow PHVT$

ii) $\tau_{VT} \rightarrow DCT \Rightarrow GDVT$

iii) $\tau_{VT} \rightarrow MelFilterbank \rightarrow DCT \Rightarrow MFGDVT$

iv) $\tau_{VT} \rightarrow Mel\ Filterbank \rightarrow Boost \rightarrow DCT \Rightarrow BMFGDVT$

Feature Extraction for ASR

Feature	TestSet A	TestSet B	TestSet C
MFCC	66.2	71.4	64.9
PLP	67.3	70.6	66.2
PNCC	71.2	72.8	71.5
MODGDF	64.3	66.4	59.5
CGDF	67.0	73.0	59.4
PS	66.0	71.2	64.6
i) PHVT	69.0	74.8	67.1
ii) GDVT	70.5	75.9	69.1
iii) MFGDVT	72.8	77.3	72.8
iv) BMFGDVT	**73.2**	**77.4**	**73.4**

i) $arg[X_{VT}] \rightarrow DCT \Rightarrow PHVT$

ii) $\tau_{VT} \rightarrow DCT \Rightarrow GDVT$

iii) $\tau_{VT} \rightarrow MelFilterbank \rightarrow DCT \Rightarrow MFGDVT$

iv) $\tau_{VT} \rightarrow Mel\ Filterbank \rightarrow Boost \rightarrow DCT \Rightarrow BMFGDVT$

* Aurora 2
* Average of 0-20 dB

Conclusion

- This talk was about phase-based source-filter deconvolution
- Separation was done using Trend/Fluctuation analysis of the phase spectrum of the minimum-phase component of speech
- Proposed method succeeds in decomposing the speech into vocal tract and excitation components
- Extracted feature from the vocal tract component of the phase shows good robustness on Aurora 2 task

That is it!

- Thanks for your attention
- Question