STATISTICAL NORMALISATION OF PHASE-BASED FEATURE REPRESENTATION FOR ROBUST SPEECH RECOGNITION

Erfan Loweimi, Jon Barker and Thomas Hain

Speech and Hearing Research Group (SPandH), University of Sheffield, Sheffield, UK

{e.loweimi1, j.p.barker, t.hain}@sheffield.ac.uk

Abstract

** Phase Spectrum is generally assumed to have a Uniform distribution

** Uniform distribution implicitly means that phase spectrum has maximum level of entropy and literally structureless/informationless

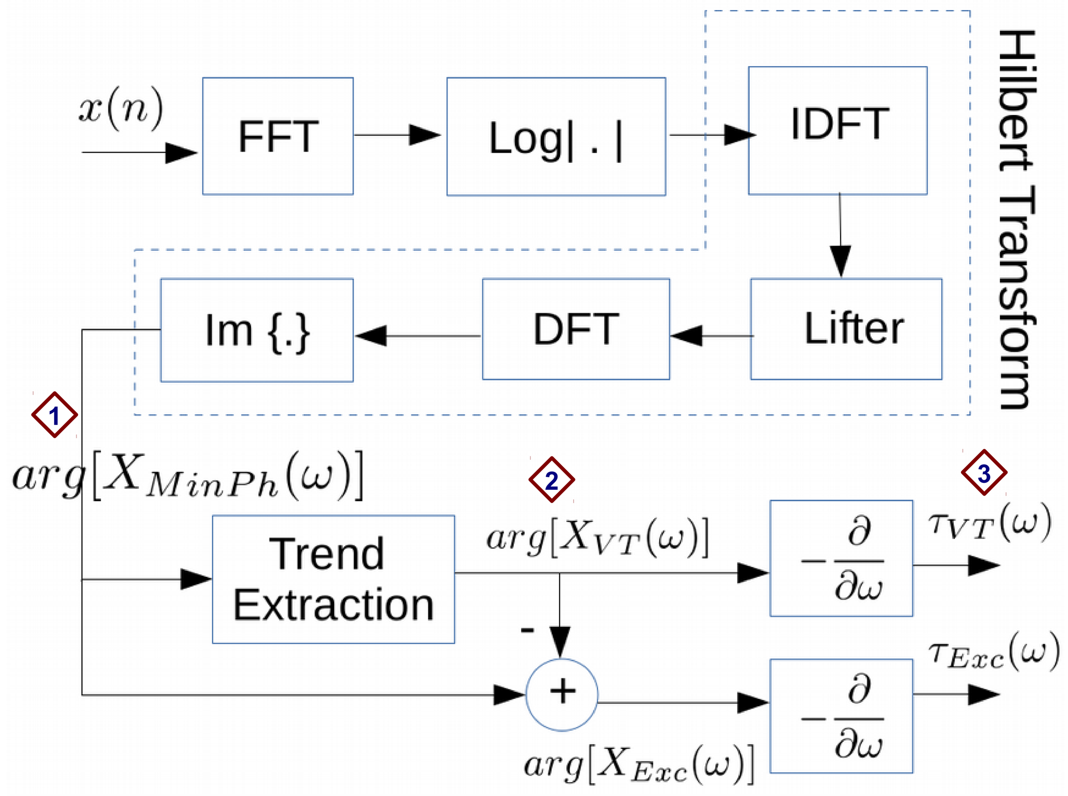
** This is paradoxical …
– Signal and its information are recoverable from the phase through phase-only signal reconstruction
– One-to-one relationship between phase and magnitude spectra necessitate both carry the same amount of information

** We show that phase spectrum, contrary to the general belief, has a bell-shaped distribution

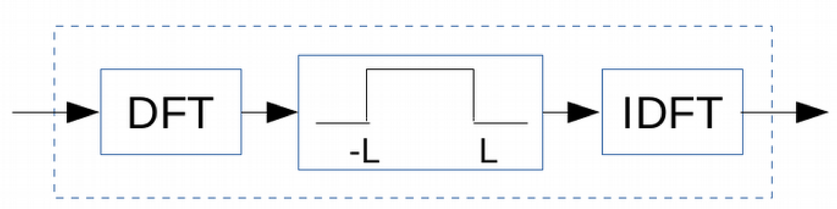
** Based on statistical behaviour of the phase-based features in clean condition, 3 normalisation schemes are applied to alleviate the effect of noise

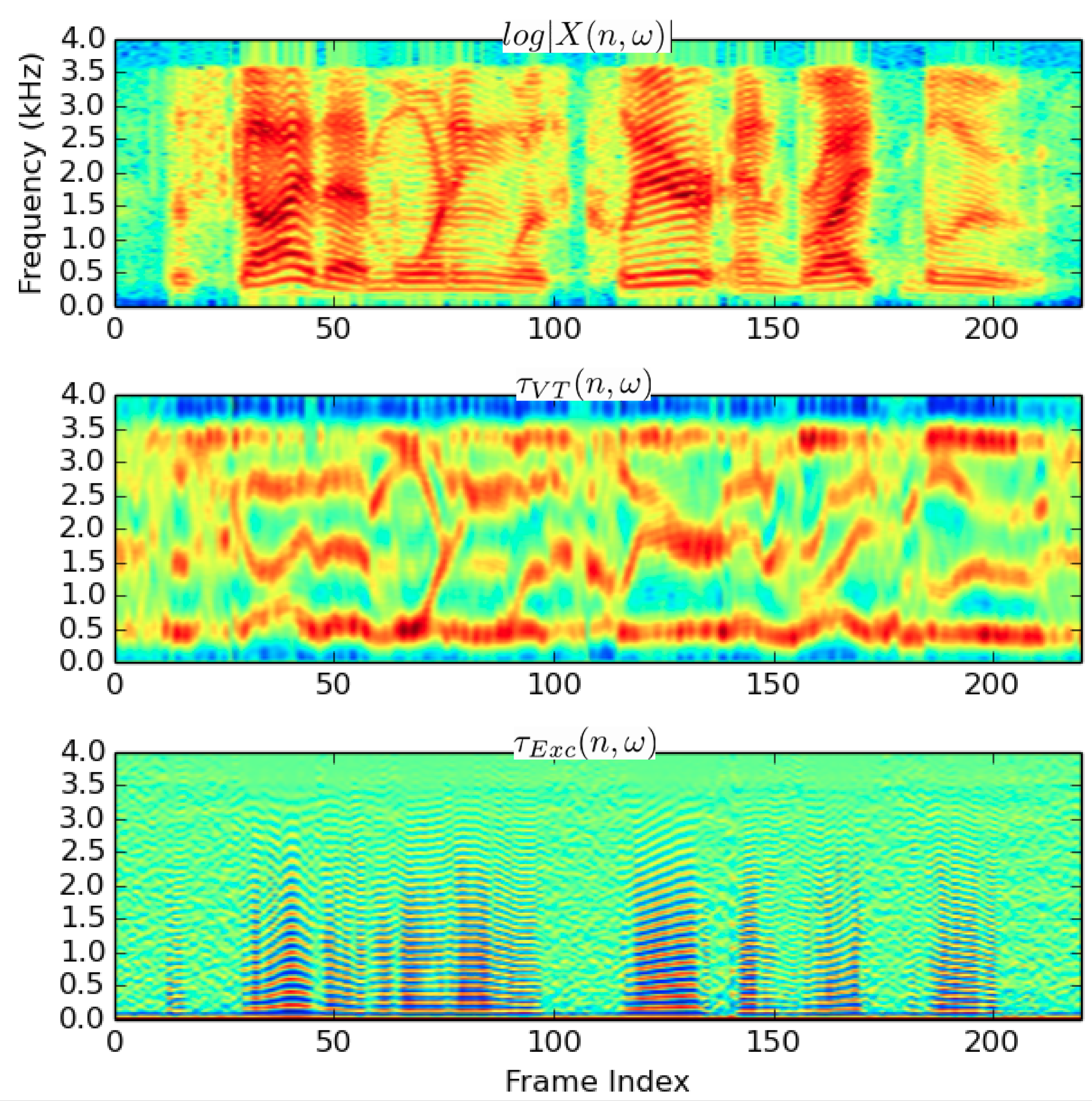
** The proposed approach returns up to 18.6% relative WER reduction compared with previous reported results [Table 1-4]

Source-Filter Separation in the Phase Domain



Trend Extraction

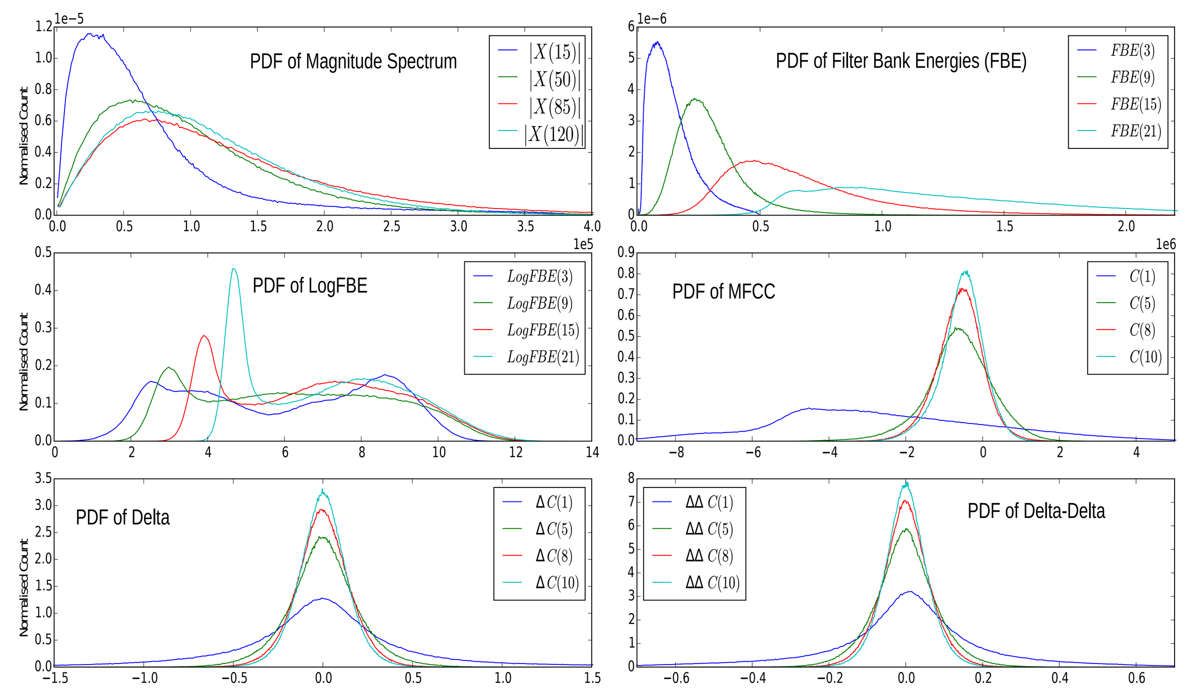




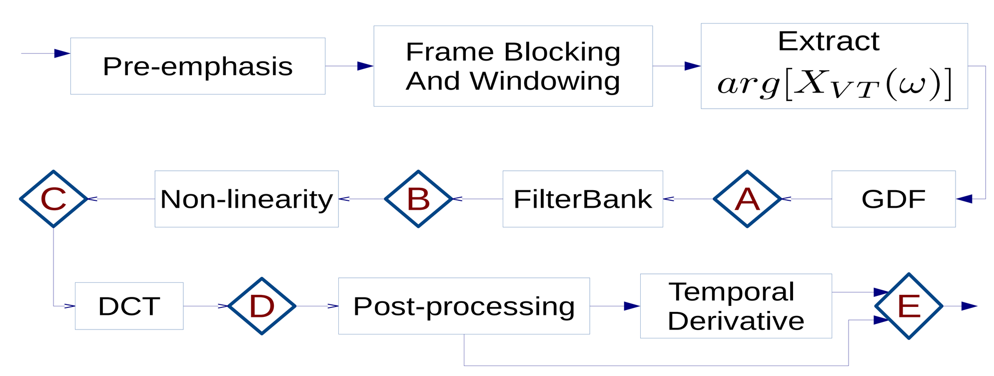
Distribution Evolution Along MFCC Pipeline

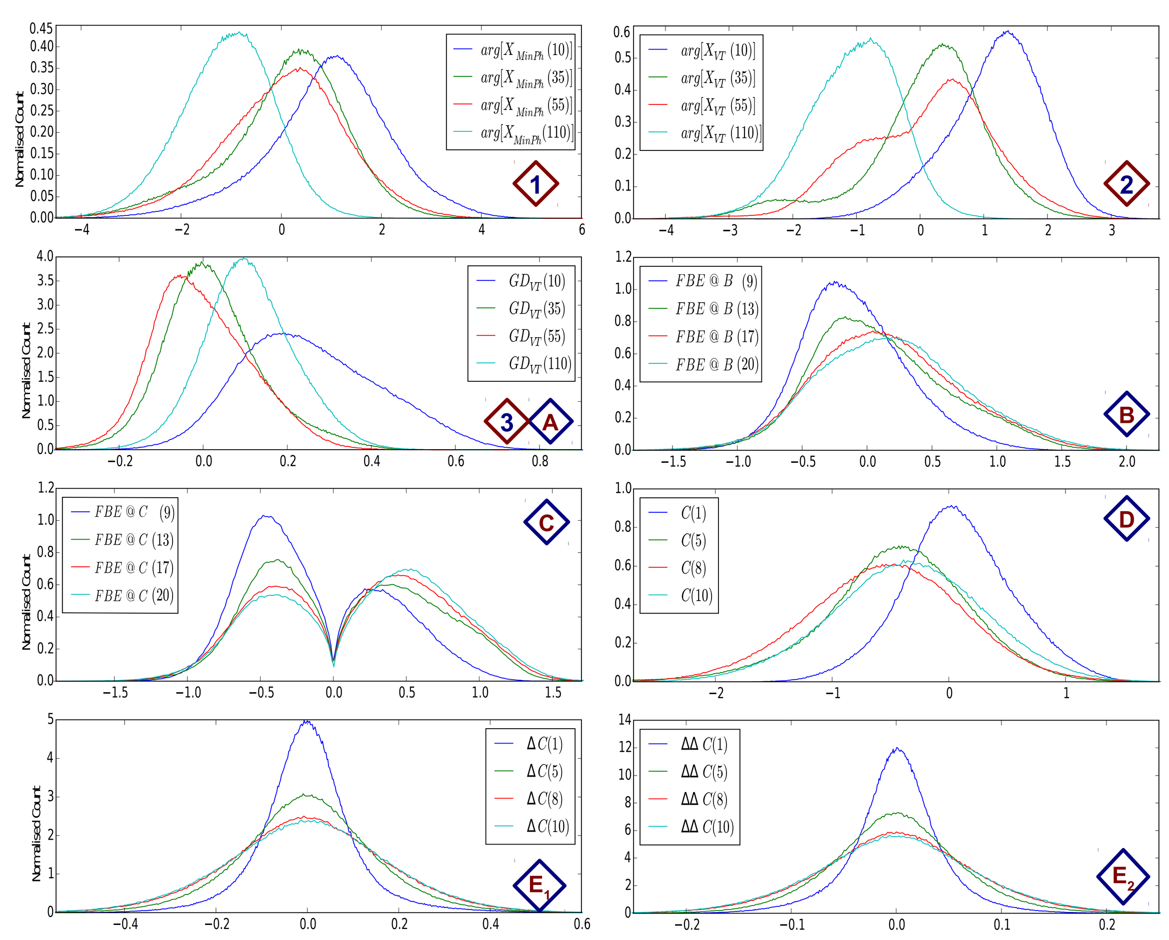
Distributions (Histograms) are computed

– using all Aurora2 training data (> 1.4 M frames)
– with suboptimal assumption that dimensions are independent (for mathematical convenience)

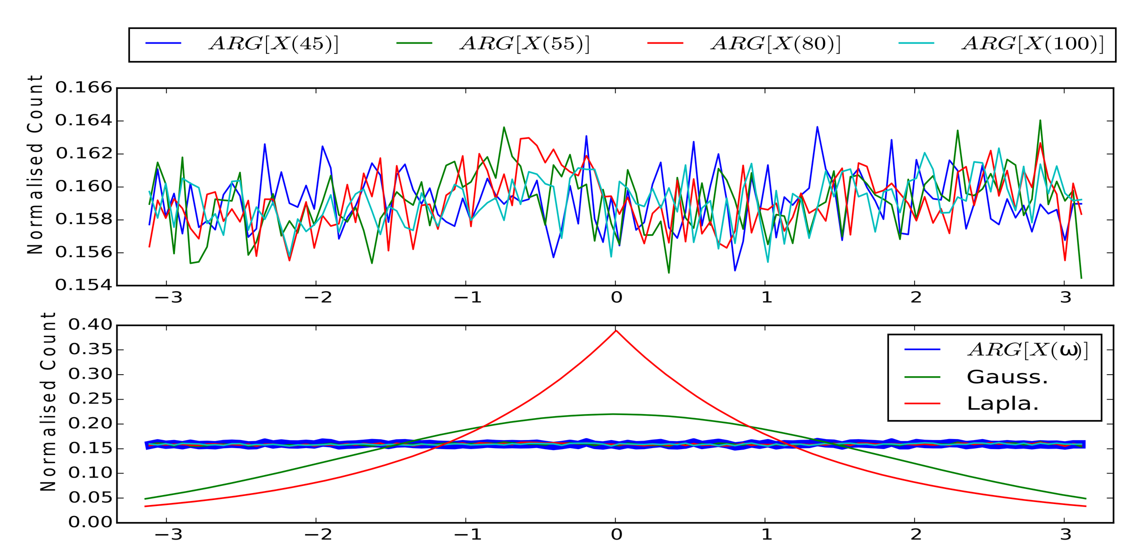


Distribution Evolution of Phase-Based Feature



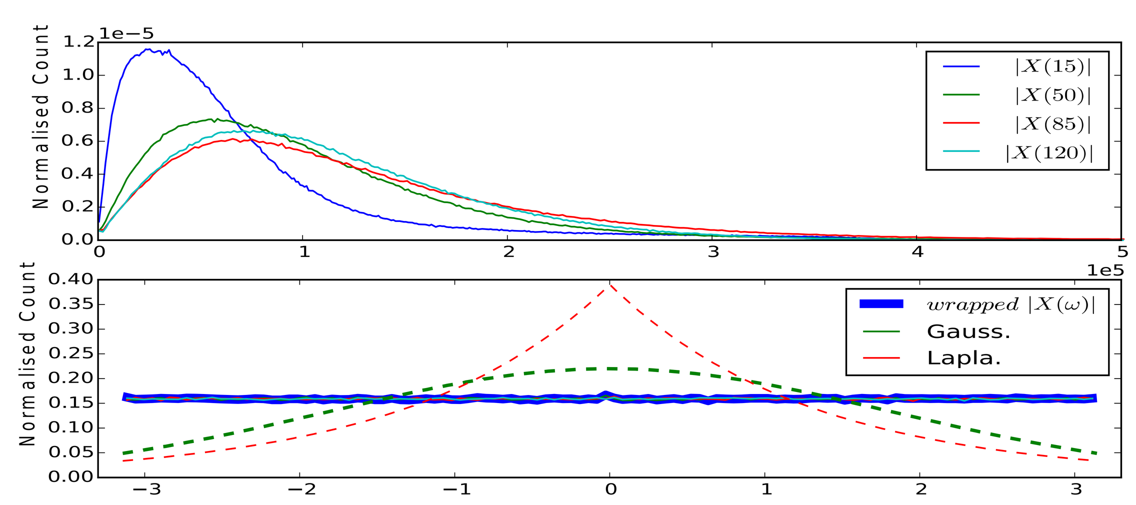


Distribution of Principle (Wrapped) Phase (ARG)

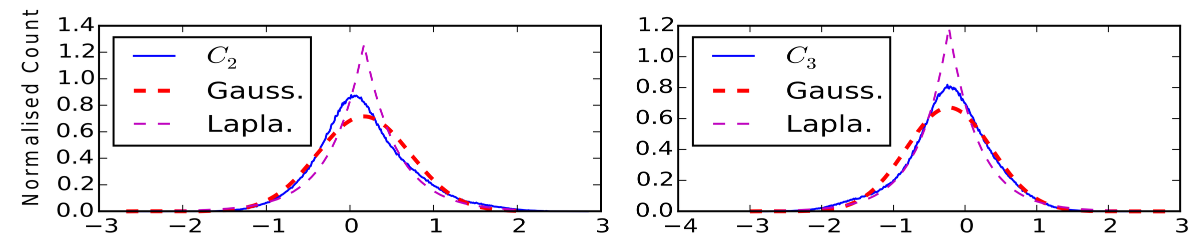


- Uniform distribution for phase spectrum is
  – paradoxical !
  – artefact of wrapping !

Distribution of Wrapped Magnitude Spectrum



Distribution of Phase-based Features



In *Clean* Condition …

– *Skewness* almost zero
– Gaussian < *Kurtosis* < Laplacian

Statistical Normalisation

Principle Equation …

$$CDF_Y(y) = CDF_X(x) \Rightarrow x = CDF_X^{-1}(CDF_Y(y))$$

– Probability Integral Transform $\Rightarrow CDF_Y(Y) \sim U(0,1)$
– Main Challenge $\Rightarrow$ Computing $CDF_X^{-1}(x)$

$$\begin{cases} Gaussianisation \;\rightarrow\; x_i = \sqrt{2}\, erf^{-1}(2z_i - 1) \\ Laplacianisation \;\rightarrow\; x_i = \begin{cases} ln(2z_i), & z_i < 0.5 \\ -ln(2-2z_i), & z_i \geq 0.5, \end{cases} \\ z_i = \frac{r_i - \beta}{N} \;,\; i = 1,\, 2,\, \dots,\, N \end{cases}$$

Recognition Results

Table 1. *Average (0-20 dB) recognition rates for Aurora-2 [23].*

Feature	TestSet A	TestSet B	TestSet C	Ave. All
MFCC	66.2	71.4	64.9	67.5
PLP	67.3	70.6	66.2	68.0
MODGDF	64.3	66.4	59.5	63.4
CGDF	67.0	73.0	59.4	66.5
PS	66.0	71.2	64.6	67.3
Baseline	**73.2**	**77.4**	**73.4**	**74.7**

– Baseline: *BMFGDVT* [Interspeech 2015]
– Normalisations are applied on both Train and Test data

Table 2. *Average accuracy after Gaussianisation at points $A-E$.*

Feature	A	B	C	Ave. All	RER(%)
Gaus-A	74.1	78.3	74.4	75.6	3.6
Gaus-B	73.0	76.0	74.1	74.4	-1.9
Gaus-C	74.0	76.7	74.9	75.2	2.0
Gaus-D	78.6	80.2	77.0	78.6	15.4
Gaus-E	79.3	81.0	77.8	**79.4**	18.6

Table 3. *Average accuracy after Laplacianisation at points $A-E$.*

Feature	A	B	C	Ave. All	RER(%)
Lap-A	74.4	78.5	74.8	75.9	4.7
Lap-B	73.9	76.7	74.8	75.1	1.6
Lap-C	74.0	76.7	75.2	75.3	2.4
Lap-D	75.5	77.5	74.0	75.7	4.0
Lap-E	77.5	79.3	75.9	**77.6**	11.5

Table 4. *Average accuracy after HEQ at points $A-E$.*

Feature	A	B	C	Ave. All	RER(%)
HEQ-A	74.0	78.0	74.9	75.6	3.5
HEQ-B	74.2	78.0	75.2	75.8	4.3
HEQ-C	74.5	78.4	75.4	76.1	5.5
HEQ-D	76.5	78.2	73.5	76.1	5.5
HEQ-E	77.0	78.7	74.9	**76.9**	8.7