THE UNIVERSITY *of* EDINBURGH

SpeechWave

Deep Scattering Spectrum (DSS)
and its Applications in ASR

Erfan Loweimi

Centre for Speech Technology Research (CSTR)
University of Edinburgh

IEEE TRANSACTIONS ON SIGNAL PROCESSING, VOL. 62, NO. 16, AUGUST 15, 2014

Deep Scattering Spectrum

Joakim Andén, *Member, IEEE*, and Stéphane Mallat, *Fellow, IEEE*

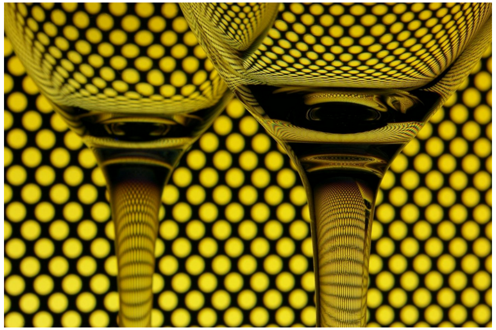
CSTR

Outline

- Deep Scattering Spectrum (DSS)
- IBM+JHU, ICASSP 2014
- IBM+JHU, INTERSPEECH 2014
- KCL+CSTR, INTERSPEECH 2020
- Wrap-up

Goal: Construct a representation ...

- ... preserves info while remains *invariant* and *stable* to variabilities within class, for example ...
  - Stable to (small) deformation, e.g. time warping
  - Invariant to geometric transformations, e.g. translation, scale

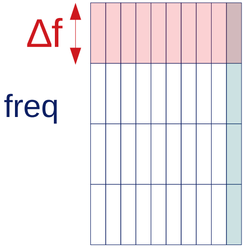


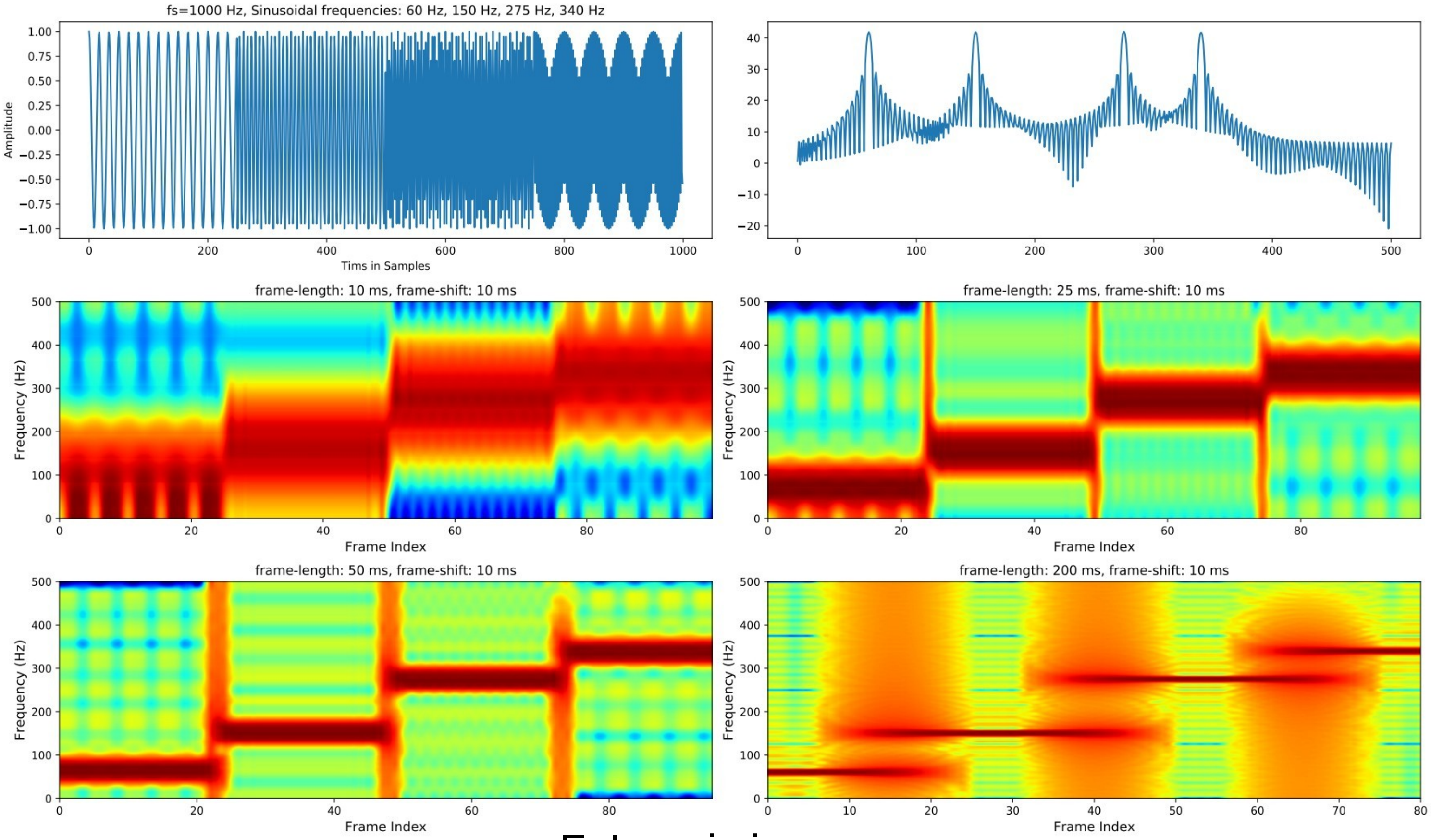
Detour

- Time-Frequency Analysis
- Wavelet Transform
- Amplitude Demodulation
- Time-warping Deformation
- Lipschitz Stability

Time-Frequency Analysis (TFA)

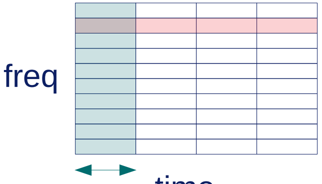
STFT, small Δt





FT

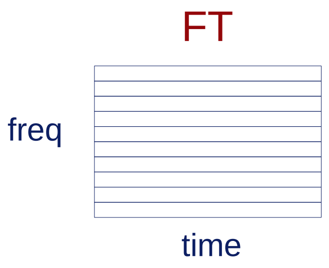
STFT, small Δf



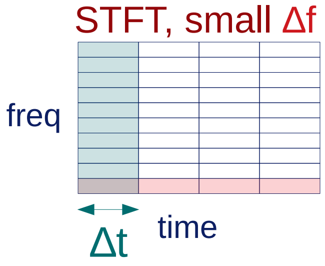
E. Loweimi

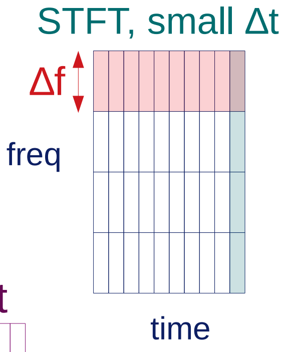
Time-Frequency Analysis (TFA)

$$\text{FT:}\; X(\omega) = \int x(t)\; e^{-j\omega t} dt$$

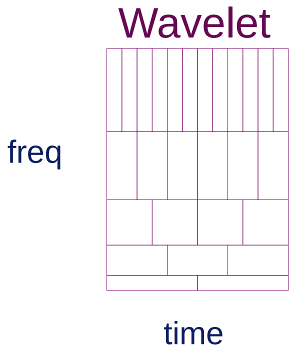


$$\text{STFT:}\; X(t,\omega) = \int x(t')\overbrace{w(t'-t)}^{\text{window}}\; e^{-j\omega t'} dt'$$





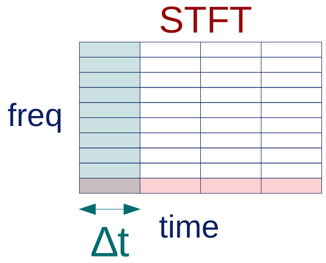
$$\text{Wavelet:}\;\; X(a,b) = \frac{1}{\sqrt{\mid a \mid}} \int x(t)\; \psi^*(\frac{t-b}{a})\; dt$$

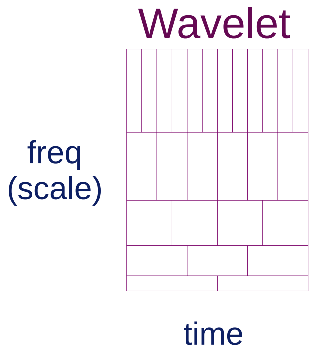


Wavelet Transform

Mother wavelet, shift, scale

$$X(a,b) = \frac{1}{\sqrt{|a|}} \int x(t)\, \underbrace{\psi^*}_{\text{Mother wavelet}}\Big(\frac{t - \overbrace{b}^{\text{shift}}}{\underbrace{a}_{\text{scale}}}\Big)\, dt$$

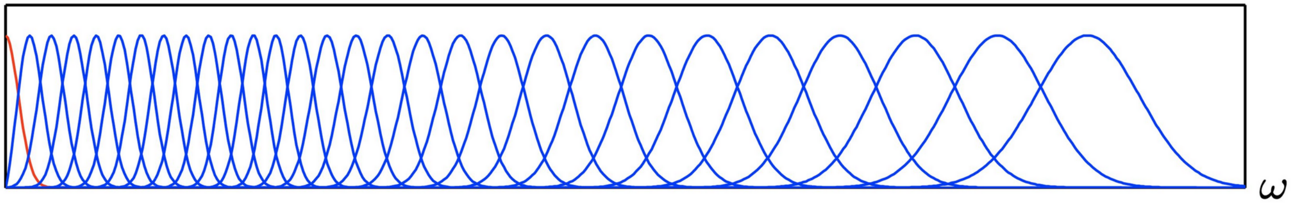




$$\text{scale} \propto \frac{1}{\text{frequency}}$$

Example:
(Complex) Morlet Wavelet

$$\psi(t) \propto \exp\Big(-\frac{t^2}{2\sigma^2}\Big)\ \exp(j2\pi f_c t)$$

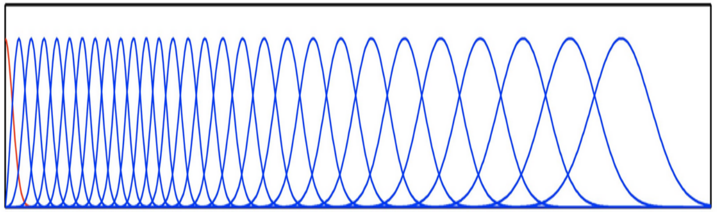


Const-Q $\leftrightarrow$ $\sigma f_c = Cte$

Wavelet Transform

- Wavelet is a filterbank, defined in time domain
- Conv. with each filter (ψ_λ) returns subband signal, $x_\lambda(t)$
- $x_\lambda(t)$ is complex; | . | → extract envelop
  - Assume $x_\lambda(t)$ is an *analytic signal*

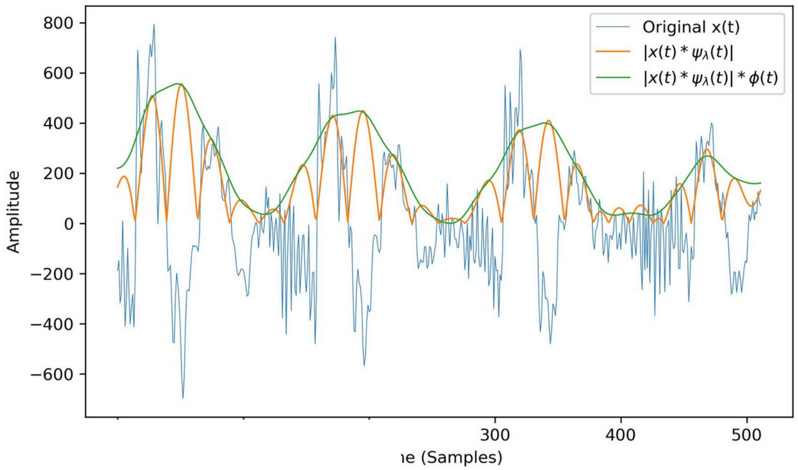
$$x_\lambda(t) \; = | \; x(t) * \psi_\lambda(t) \; |$$



$$x_{\text{analytic}}(t) = x(t) + j\mathcal{H}\{x(t)\}$$
$$|\; x_{\text{analytic}}(t) \;| = \text{Envelope of } x(t)$$

Amplitude Demodulation

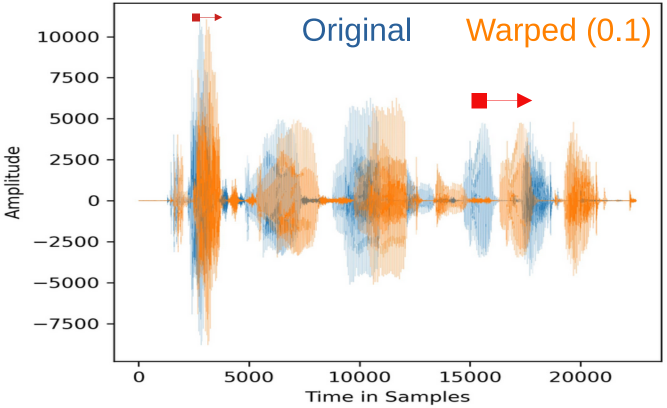
$$\left|x(t) * \psi_\lambda(t)\right| * \phi(t)$$

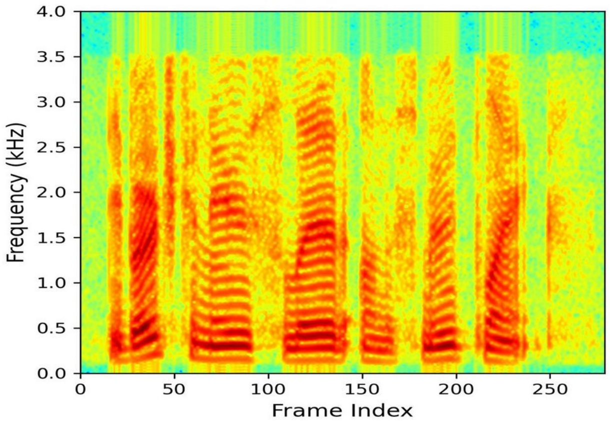


- $\phi(t)$: Low-pass filtering
- $|\ .\ | * \phi(t)$: Extract Envelop (amplitude demodulation)
- $x(t) * \psi_\lambda(t)$: Extract subband signal
- $|x(t) * \psi_\lambda(t)| * \phi(t)$: Extract envelop of subband signal

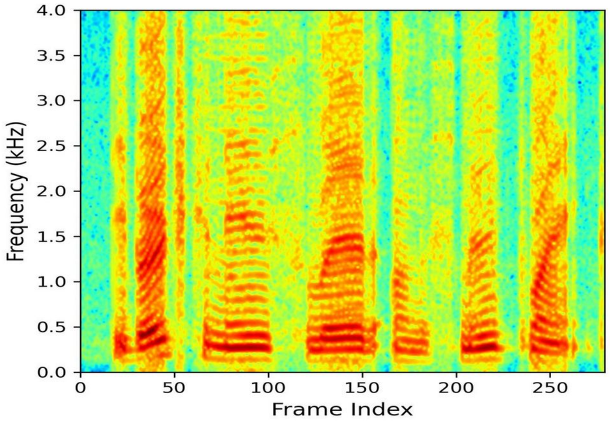
Time-warping Deformation (TWD)

- Variable time-shift
  - Definition: $x(t) = x_{\tau}(t - \tau(t))$
  - Example: $\tau(t) = \varepsilon\, t$





Original



Warped

ε=0.1

ε=0.2

Lipschitz Stability

- Stability: small deformation in **x** ==>> small change in **Φ**(**x**)
  - Deformation size measured by $\sup_t |\nabla \tau(t)|$
  - Change size → Euclidean distance
- *Φ(x)* is Lipschitz stable to deformation $x_\tau(t)$ if a C > 0 exists s.t.

$$\forall\, \tau,\ ||\Phi(x) - \Phi(x_\tau)|| \leq C \ \sup_t |\nabla \tau(t)|\ ||x||$$

- The lower the C, the higher the stability

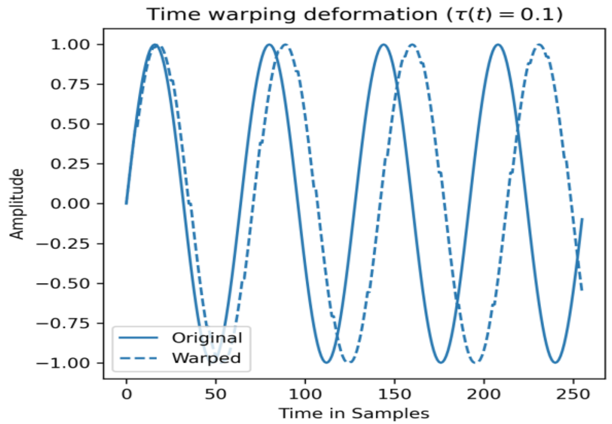
Spectrogram (|X(t,ω)|)

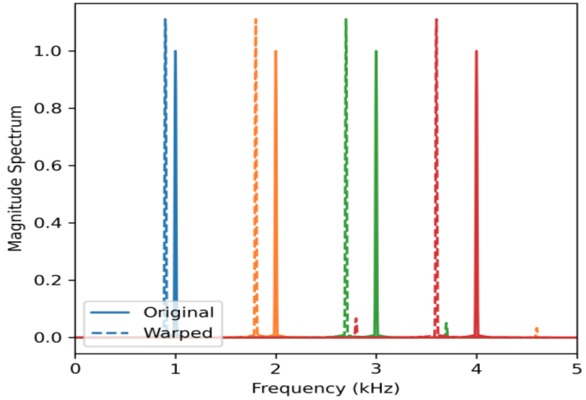
- Invariant to time-shift (c) [:-)|
- Unstable to TWD [:-(|
  - Larger *ω* → Larger *Δω* => *No C!*

$$x_c(t) = x(t-c) \xrightarrow{\mathcal{F}} e^{-j\omega_k n_0} X(\omega) \xrightarrow{|.|} |X(\omega)|$$

$$x_\tau(t) = x(t-\tau(t)) = x(t-\epsilon t)$$

$$x_\tau(t) \xrightarrow{\mathcal{F}} X_\tau(\omega) = \frac{1}{1-\epsilon} X(\frac{\omega}{1-\epsilon}) \xrightarrow{|.|} \approx |X_\tau(\omega)|$$





Mel-Spectrogram

- $H(\omega; \lambda_i)$: frequency response of i^{th} filter (λ_i = centre frequency)
- Role: frequency **averaging** + subsampling ≈ avg pooling
  - Makes $Mx(t; \lambda_i)$ Lipschitz stability (unlike $|X(t,\omega)|$) [:-)]
  - Brings about irreversible information loss [:-(]

$$Mx(t, \lambda_i) = \int_\omega |X(t,\omega)|^2 |H(\omega; \lambda_i)|^2 d\omega$$

$$= \int_{t'} |x(t,t') * h(t'; \lambda_i)|^2 dt'$$

Plancherel Theorem

Mel-Spectrogram

- $H(\omega; \lambda_i)$: frequency response of i^{th} filter (λ_i = centre frequency)
- Role: frequency **averaging** + subsampling ≈ avg pooling
  - Makes $Mx(t; \lambda_i)$ Lipschitz stable (unlike $|X(t,\omega)|$) [:-)|
  - Brings about irreversible information loss [:-(|

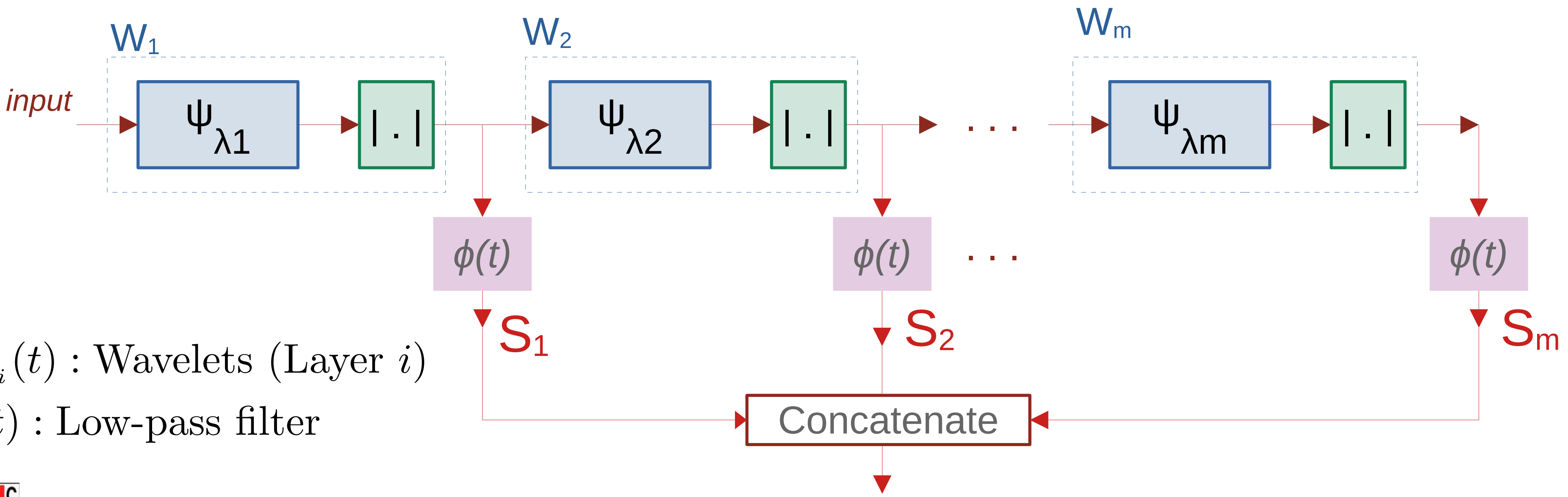
$$Mx(t, \lambda_i) = \int_\omega |X(t,\omega)|^2 |H(\omega; \lambda_i)|^2 \, d\omega$$
$$\approx \boxed{|x(t) * h(t; \lambda_i)|^2 * \phi^2(t)}$$

Proof in Appendix **A**


END
DETOUR

Scattering Transform (1)

- A cascade of Wavelet transforms (linear) and modulus (non-linear)



$\psi_{\lambda_i}(t)$: Wavelets (Layer i)

$\phi(t)$: Low-pass filter

Scattering Transform (2)

- A cascade of Wavelet ($\Psi_\lambda(t)$) transforms and modulus (| . |)

0th order $$\mathrm{S}_0\mathrm{x}(t) = x(t) * \phi(t)$$

1st order $$\mathrm{S}_1\mathrm{x}(t,\lambda_1) = |x(t) * \psi_{\lambda_1}(t)| * \phi(t)$$

2nd order $$\mathrm{S}_2\mathrm{x}(t,\lambda_1,\lambda_2) = ||\, x(t) * \psi_{\lambda_1}(t)| * \psi_{\lambda_2}(t)| \; * \; \phi(t)$$

.
.
.

$$\vdots$$

Mth order $$\mathrm{S}_m\mathrm{x}(t,\lambda_1,\cdots,\lambda_m) = |\, \cdots |x(t) * \psi_{\lambda_1}(t)| * \cdots \,| * \psi_{\lambda_m}(t)| \; * \; \phi(t)$$

Scattering Transform (3)

- A cascade of Wavelet ($\Psi_\lambda(t)$) transforms and modulus
- $\phi(t)$: low-pass filter → averaging → stability + information loss

0th order $$\mathrm{S}_0\mathrm{x}(t) = x(t) * \phi(t)$$

1st order $$\mathrm{S}_1\mathrm{x}(t, \lambda_1) = |x(t) * \psi_{\lambda_1}(t)| * \phi(t)$$

2nd order $$\mathrm{S}_2\mathrm{x}(t, \lambda_1, \lambda_2) = || \, x(t) * \psi_{\lambda_1}(t)| * \psi_{\lambda_2}(t)| \; * \; \phi(t)$$

$$\vdots$$

Mth order $$\mathrm{S}_m\mathrm{x}(t, \lambda_1, \cdots, \lambda_m) = | \cdots |x(t) * \psi_{\lambda_1}(t)| * \cdots | * \psi_{\lambda_m}(t)| \; * \; \phi(t)$$

Role of Scattering Coefficients

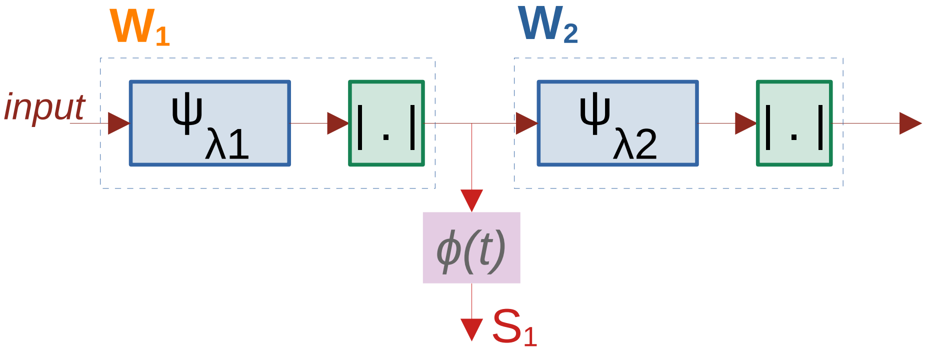
- First order scattering coef. (S_1) ≡ filterbank energies
- S_m aims at compensating for lost info in S_{m-1}
- Information loss … due to low-pass filtering ...
  - Fast temporal transients (high freq.) info, e.g. attack, is lost!

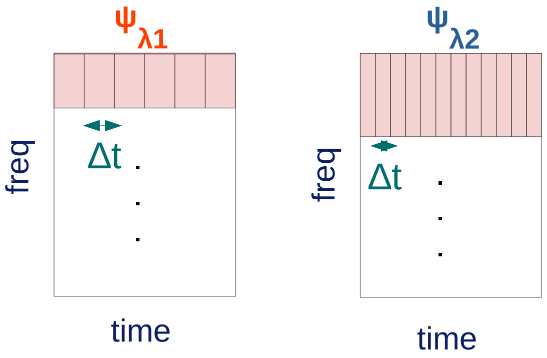
Role of Scattering Coefficients

- First order scattering coef. (S_1) ≡ filterbank energies
- S_m aims at compensating for lost info in S_{m-1}
- Information loss … due to low-pass filtering ...
  - Fast temporal transients (high freq.) info, e.g. attack, is lost!
- **Solution**: Another transform with a higher *time resolution*
  - … should better localise the transients in time

Role of high order Scattering Coef.

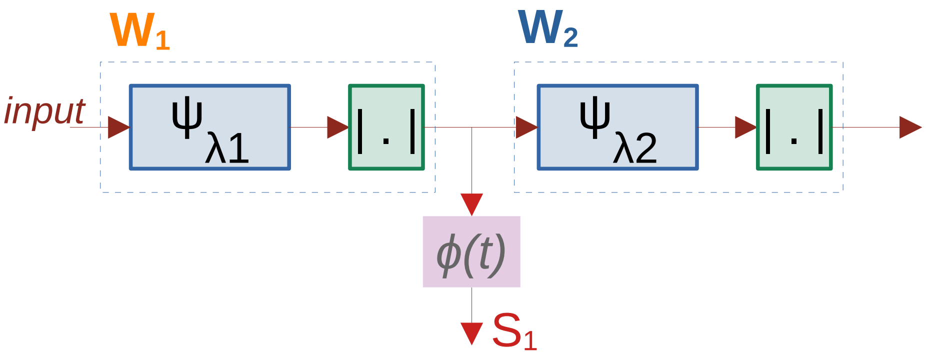
- $\Psi_{\lambda 2}$ should have a smaller Δt than $\Psi_{\lambda 1}$
  - $\Psi_{\lambda 2}$'s filters should be narrower in time domain
    - wider in frequency domain

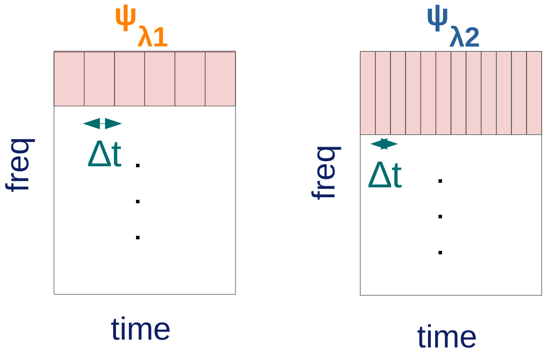




Role of high order Scattering Coef.

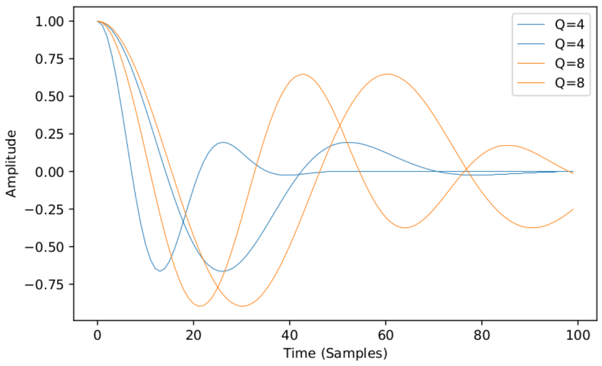
- $\Psi_{\lambda 2}$ should have a smaller Δt than $\Psi_{\lambda 1}$
  - $\Psi_{\lambda 2}$'s filters should be narrower in time domain (wider in Hz)
- Ψ_{λ} is in a constant-**Q** filterbank (Q ≡ knob)
  - $Q_1 > Q_2$ or $Q_1 < Q_2$?

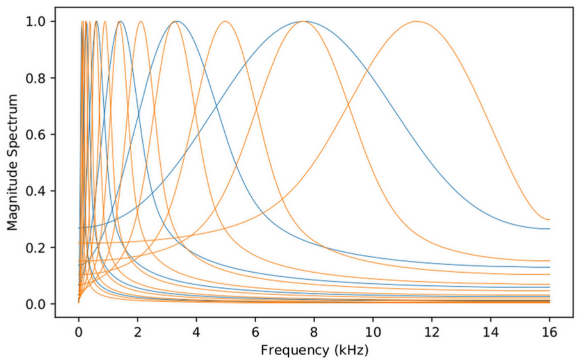


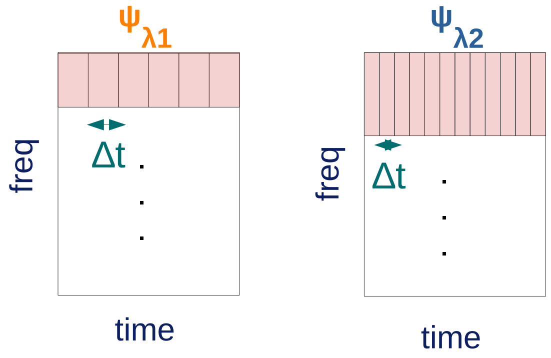


Role of high order Scattering Coef.

- $\Psi_{\lambda 2}$ should have a smaller Δt than $\Psi_{\lambda 1}$
  - $\Psi_{\lambda 2}$'s filters should be narrower in time domain
- Smaller Q $\rightarrow$ filters wider in freq domain $\rightarrow$ narrower in time

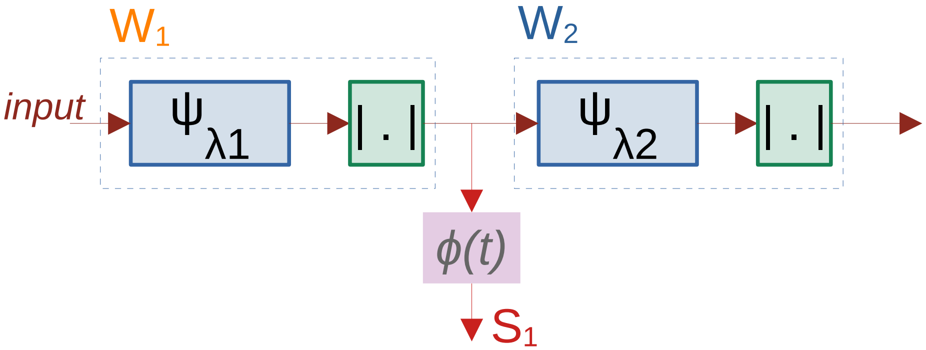


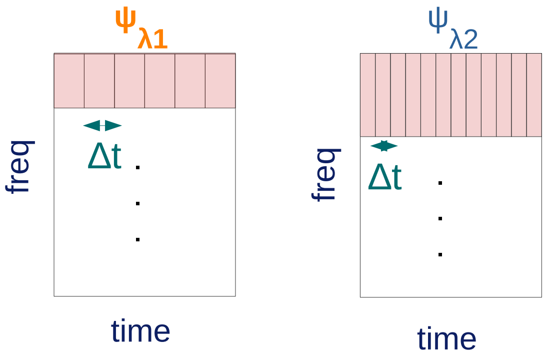




Role of high order Scattering Coef.

- $\Psi_{\lambda 2}$ should have a smaller Δt than $\Psi_{\lambda 1}$
  – $\Psi_{\lambda 2}$'s filters should be narrower in time domain
- Ψ_{λ} is in a constant-**Q** filterbank ($Q \equiv$ knob)
  ✔ Q2 < Q1





Sparsity of Higher Order Coef

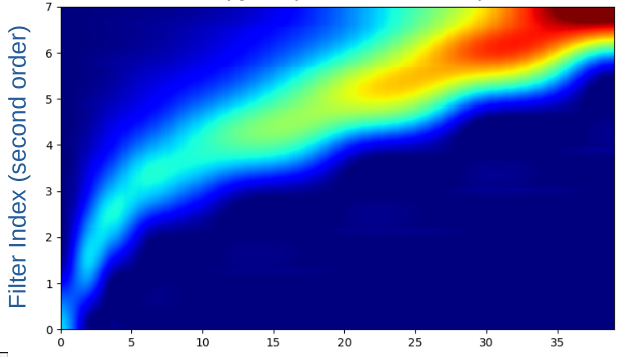
- Higher order coef are sparse (mostly zero)
- Non-zero if $\psi_{\lambda 1}$ and $\psi_{\lambda 2}$ overlap
- Only compute *non-negligible* coefficients ...

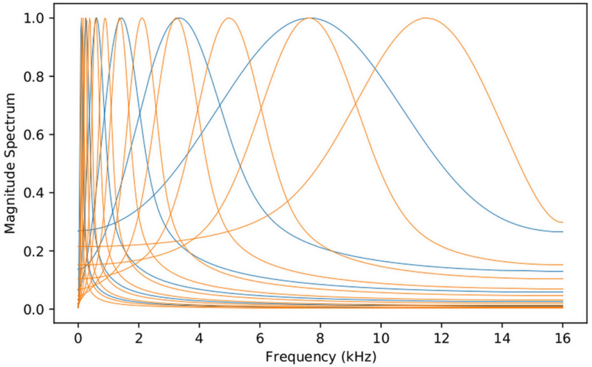
$$\lambda_2 \leq \max(\lambda_1/Q_1, 2\pi/T)$$

Centre frequency (rad/s) of the second order filters

Centre frequency (rad/s) and quality of the first order filters

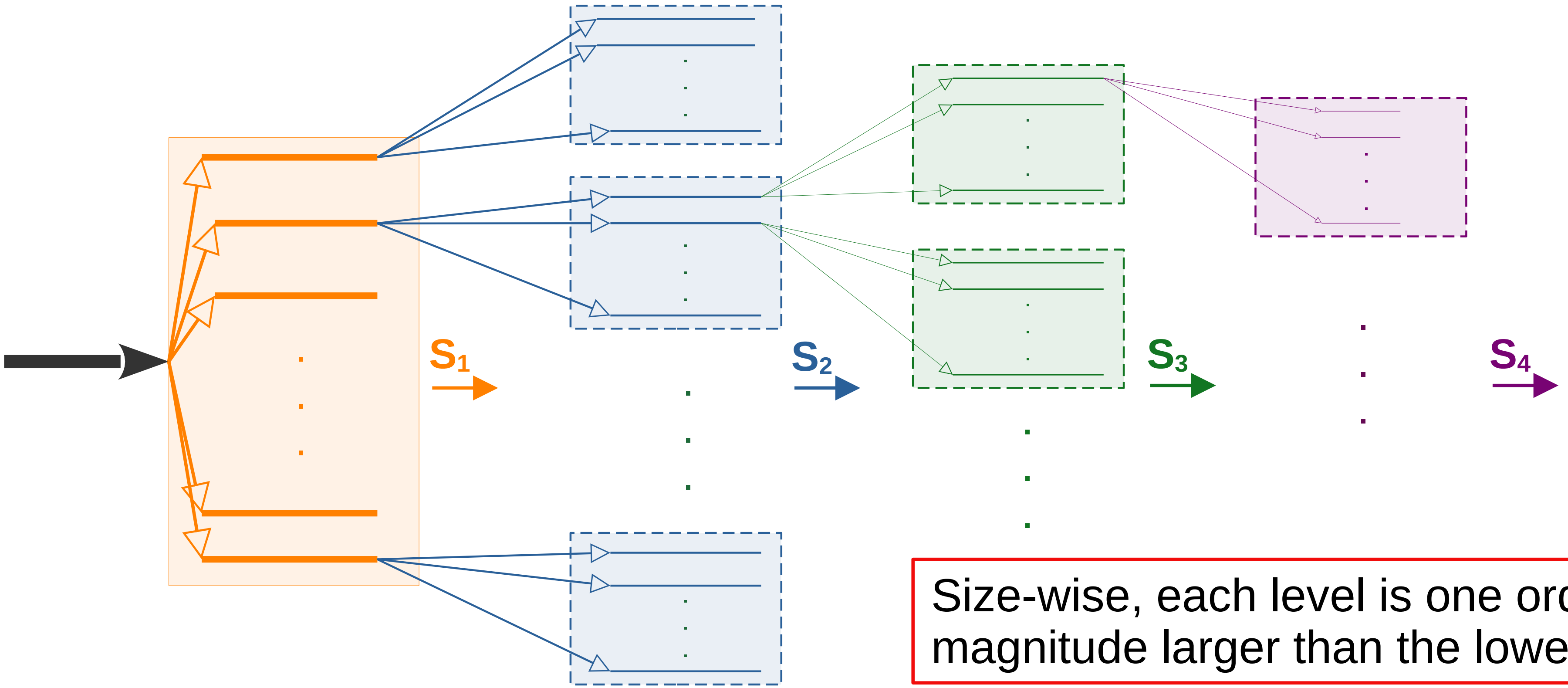
Window Length





E. Loweimi

Dimension of Scattering Coef.



Size-wise, each level is one order of magnitude larger than the lower level!

Energy (?) of Scattering Coef.

- For 25ms signal decomposition ...
  - 94.5% of energy is in S_1, ~ 4.8% in S_2
- By frame extension energy of high order Coef. increases
  - Not useful for speech, but may be music

Averaged $\| S_m \mathrm{x} \|^2 / \| \mathrm{x} \|^2$

T	$m=0$	$m=1$	$m=2$	$m=3$
23 ms	0.0%	94.5%	4.8%	0.2%
93 ms	0.0%	68.0%	29.0%	1.9%
370 ms	0.0%	34.9%	53.3%	11.6%
1.5 s	0.0%	27.7%	56.1%	24.7%

Normalising Scattering Coef.

- Normalise order *m* with order *m-1*
- Goal: improve invariability, e.g. to channel distortion

$$S_1(t,\lambda_1) = \frac{S_1(t,\lambda_1)}{S_0(t,\lambda_1) + \boxed{\epsilon}} \qquad S_2(t,\lambda_1,\lambda_2) = \frac{S_2(t,\lambda_1,\lambda_2)}{S_1(t,\lambda_1) + \boxed{\epsilon}}$$

Silence detection threshold
(to avoid x/0)

$$S_m(t,\lambda_1,\cdots,\lambda_m) = \frac{S_m(t,\lambda_1,\cdots,\lambda_m)}{S_{m-l}(t,\lambda_1,\cdots,\lambda_{m-1}) + \boxed{\epsilon}}$$

Normalising Scattering Coef.

- Normalise order *m* with order *m-1*
- Goal: improve invariability, e.g. to channel distortion

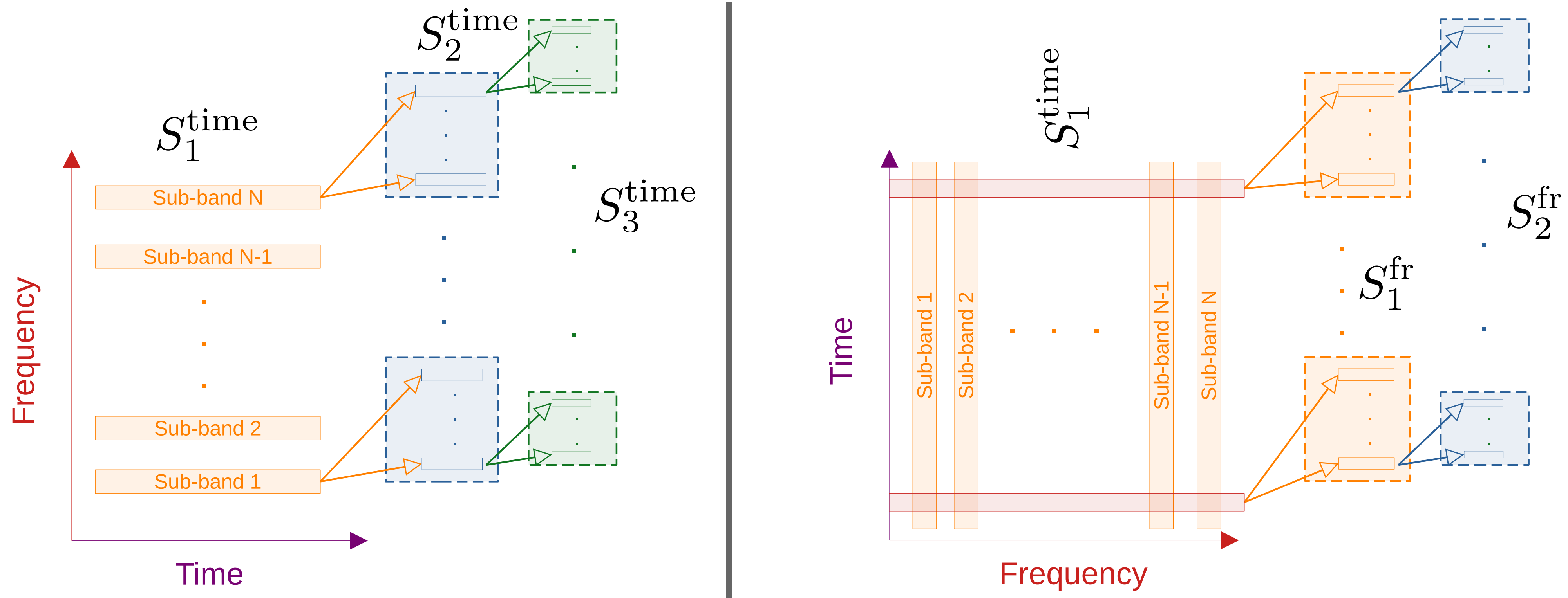
$$S_m(t, \lambda_1, \cdots, \lambda_m) = \frac{S_m(t, \lambda_1, \cdots, \lambda_m)}{S_{m-l}(t, \lambda_1, \cdots, \lambda_{m-1}) + \epsilon}$$

$$h(t) * \psi_\lambda(t) \approx |H(\omega = \lambda)| \, \psi_\lambda(t)$$

$$|\, (x(t) * h(t)) * \psi_\lambda(t) \,| \approx |H(\omega = \lambda)| \, |x(t) * \psi_\lambda(t)|$$

Holds only when H(ω) is approximately constant over support of *ψ(ω;λ)*

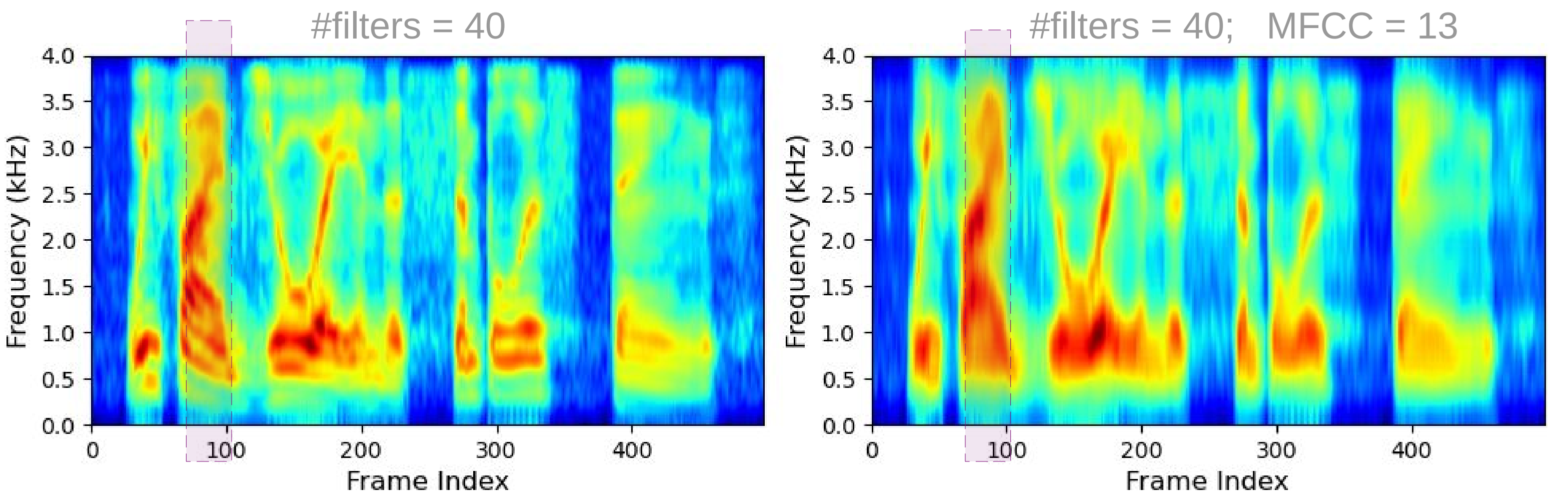
Frequency Scattering (1)



Time Scattering | **Frequency** Scattering

Frequency Scattering (2)

- Similar to freq. avg. by setting higher order MFCCs to 0
- Provides stability to *frequency transposition*



Frequency Scattering (2)

- Similar to freq. avg. by setting higher order MFCCs to 0
- Provides stability to *frequency transposition*
- Only the first-order is used, with small Q (e.g. Q=1)
- Filters are centred at ***quefrency*** λ

$$S^{\mathrm{fr}} z(\gamma, \lambda_q) = |z(\gamma) * \psi_{\lambda_q}(\gamma)| * \phi^{\mathrm{fr}}(\gamma)$$

$$\gamma = \log_2(\lambda)$$

Experimental Results

- Second order helps
  - Especially for music (Y?)
- Third order may slightly help
  - Costly because of dimension
- Freq. scattering helps

Representations	GTZAN	TIMIT
Δ-MFCC (T = 23 ms)	20.2 ± 5.4	18.5
Δ-MFCC (T = 740 ms)	18.0 ± 4.2	60.5
State of the art (excluding scattering)	9.4 ± 3.1 [8]	16.7 [43]
	$T = 740$ ms	$T = 32$ ms
Time Scat., $l = 1$	19.1 ± 4.5	19.0
Time Scat., $l = 2$	10.7 ± 3.1	17.3
Time Scat., $l = 3$	10.6 ± 2.5	18.1
Time & Freq. Scat., $l = 2$	9.3 ± 2.4	16.6
Adapt Q_1, Time & Freq. Scat., $l = 2$	8.6 ± 2.2	15.9

order ←

* **GTZAN**: Music Genre Classification
* TIMIT: phone classification
* Classifier: SVM with Gaussian Kernel
* Adapt → multi-resolution: Q=1, 8

Sturm, 2012, "An Analysis of the GTZAN Music Genre Dataset"
"… 5% … exact duplicates, 10.8% is mislabelled …"

Properties of Scattering Transform

- Similar to CNNs (hierarchical) but involves no learning
  - Learns a general (not task-specific) representation; interpretable
- Translation-invariant, stable to deformation, preserves info
- Some similarities to physiological models (cochlea, const-Q)
- Energy conservative and contractive mapping
- Has approximate and non-trivial inverse transformation
- Poorer frequency resolution than STFT

DEEP SCATTERING SPECTRUM WITH DEEP NEURAL NETWORKS

Vijayaditya Peddinti[†*], *Tara N. Sainath*[‡], *Shay Maymon*[‡]
Bhuvana Ramabhadran[‡], *David Nahamoo*[‡], *Vaibhava Goel*[‡]

†Center for Language and Speech processing, Johns Hopkins University, MD 21218, USA
‡ IBM T.J. Watson Research Center, Yorktown Heights, NY 10598, USA
vijay.p@jhu.edu, {tsainath,maymon,bhuvana,nahamoo,vgoel}@us.ibm.com

This paper investigates ...

- Usefulness of ...
  - DSS for TIMIT phone recognition
  - Multi-resolution DSS

- Optimal architecture for ...
  - Processing S_1 and S_2, simultaneously
  - Multi-resolution DSS

Experimental Setup

- Task: TIMIT phone recognition
- Baseline: 40-dim log-mel fbank + Δ + ΔΔ
- DNN: 2 x CNN (256 filters) → 3 x FC (1024)
- Output/Target: CI (147) and CD (2400)
- **MVN** for log-Mel and S_1; **MN** for S_2
  - Scatter transfer operator act like var-norm (?)
- **Delta** only for log-mel and S_1; not S_2 [not beneficial]

Experimental Results – TIMIT

- PERs of log-Mel and S_1 r similar
  - TIMIT, **0.3**, statistically significant?
- Using S_2 may help, but NOT consistently!
  - Why? Functionality overlap …
    - Δ and S_2? $\Delta\Delta$ and S_3?
- ReLU and Regularisation help

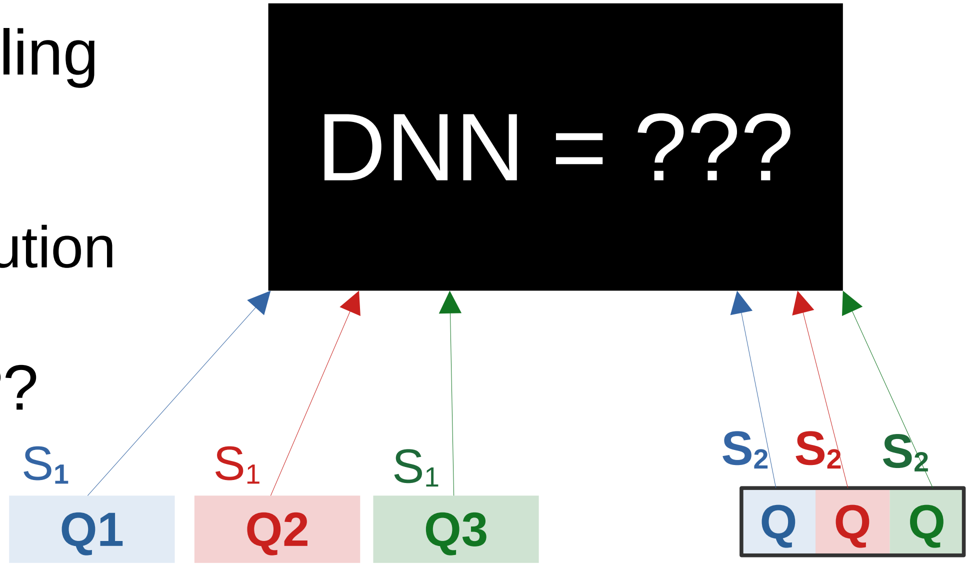
Feature	PER	
	CI	CD
$logmel + \Delta + \Delta\Delta$	19.3	18.7
$S_1 x(t, \lambda_1) + \Delta + \Delta\Delta$	19.0	18.7

Non-linearity	$S_1 + \Delta + \Delta\Delta$	$S_1 + \Delta + \Delta\Delta + S_2$
Sigmoid	21.3	20.9
ReLU	20.0	20.3
ReLU+regularization	19.0	18.8

* CI: context-independent (147)
* CD: context-dependent (2400)
* Regularisation: MaxNorm and Dropout

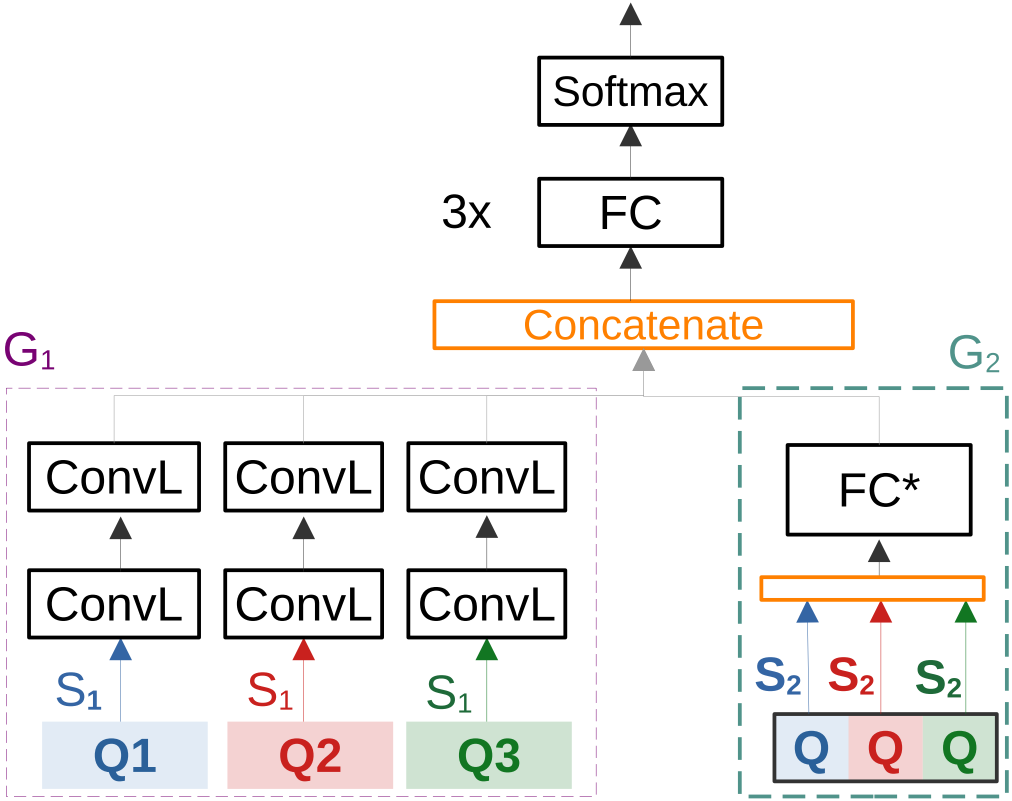
Multi-Resolution Approach

- Use multiple filterbanks with various Qs
  - ONLY for S_1; $S_2 \leftrightarrow$ *always* Q=1
- Advantage: complementary modelling
  - Small Q $\rightarrow$ better time resolution
  - Large Q $\rightarrow$ better frequency resolution
- Optimal architecture to combine???



Architecture for Multi-Resolution

- Multi-resolution ≡ Various Qs
- Process S_1 with (2x) ConvL
- Process S_2 with FC*
  - S_2 is sparse + Limited local corr
  - Not optimal for ConvL
    - Too short filters



Multi-Resolution -- TIMIT -- CI

- Multi-resolution helps!
- Multi-resolution for S_1 (G_1) is more helpful than S_2
  - 0.6 vs 0.2
- Optimal width for FC* is 128

Feature Stream	PER
S_1+Δ+ΔΔ	19.0
G_1+Δ+ΔΔ	18.4
S_1+Δ+ΔΔ+S_2	18.8
G_1+Δ+ΔΔ +G_2+1024 HU	19.1
G_1+Δ+ΔΔ +G_2+256 HU	18.7
G_1+Δ+ΔΔ +G_2+128 HU	**18.2**
G_1+Δ+ΔΔ +G_2+ 64 HU	18.6

* G_1: multi-resolution S_1
* G_2: multi-resolution S_2
* HU: #hidden units of FC*

Multi-Resolution -- TIMIT -- CD

- Using S_2 helps
  - PER: 18.7 → 17.9 [0.8]
  - For CI: 19.0 → 18.8 [0.2]
- Multi-Resolution helps
  - PER: 17.9 → 17.4 [0.5]
  - For CI: 18.8 → 18.2 [0.6]

Feature Stream	PER
S_1+Δ+ΔΔ	18.7
S_1+Δ+ΔΔ+S_2 128 HU	17.9
G_1+Δ+ΔΔ +G_2+128 HU	**17.4**

* G_1: multi-resolution S_1
* G_2: multi-resolution S_2
* HU: #hidden units of FC*

Deep Scattering Spectra with Deep Neural Networks for LVCSR Tasks

Tara N. Sainath[1], *Vijayaditya Peddinti*[2], *Brian Kingsbury*[1],
Petr Fousek[1], *Bhuvana Ramabhadran*[1], *David Nahamoo*[1]

[1]IBM T. J. Watson Research Center, Yorktown Heights, NY 10598, U.S.A
[2]Center for Language and Speech Processing, Johns Hopkins University, MD 21218, U.S.A
tsainath@us.ibm.com, vijay.p@jhu.edu, bedk@us.ibm.com,
petr_fousek@cz.ibm.com, {bhuvana, nahamoo}@us.ibm.com

interspeech 2014 singapore
Celebrating the Diversity of Spoken Languages

14-18 September 2014, Singapore

This paper investigates ...

- LVCSR (BN: 50h; BN: 430h)
- Multi-resolution + frequency scattering effect
- Dimensionality reduction
- Speaker adaptation
- Sequence training

Experimental Results

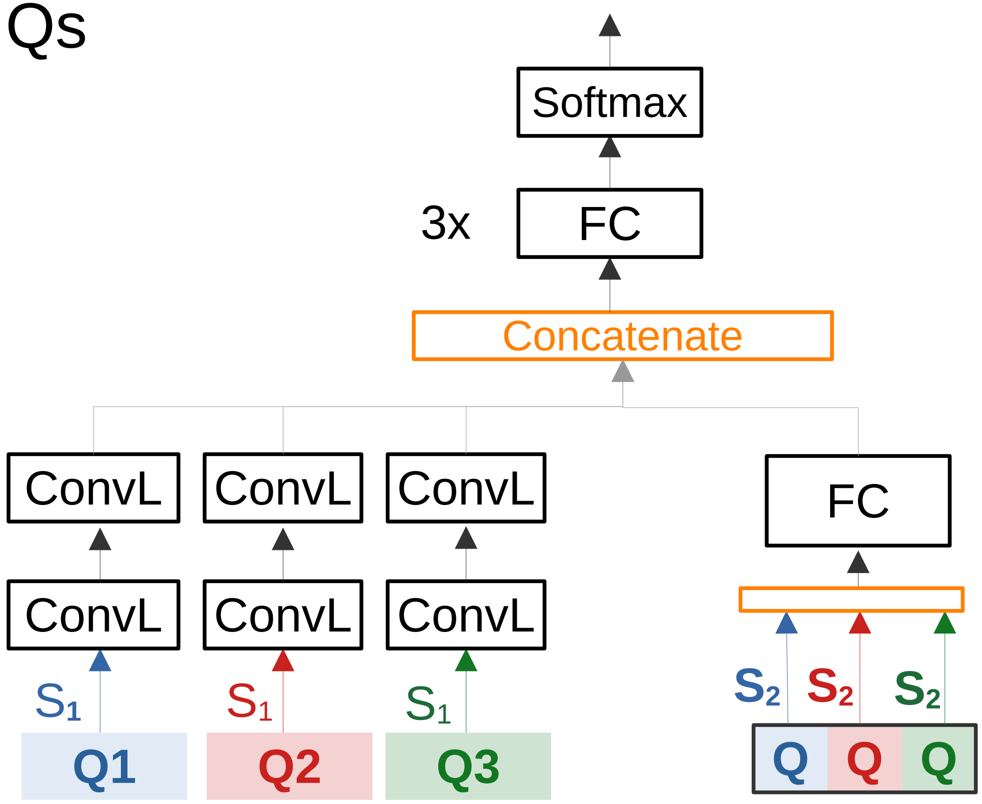
- $S_1(+S_2)$ is comparable to log-mel!
- S_2 slightly helps!
  - WER: 16.0 → 15.9
- Frequency scattering helps!
  - WER: 15.9 → 15.5
- Gain carries over to larger tasks

English Broadcast News, 50h

Feature	WER
log-mel baseline	15.9
S_1, time	16.0
S_1+S_2, time	15.9
S_1+S_2, time+frequency	**15.5**

Multi-Resolution Approach

- Multiple filterbanks with different Qs
- Various Qs ONLY for S_1
  - For S_2, always Q=1
- S_1 modelled by ConvL
- S_2 modelled by FC
  - S_2 is sparse; Limited local corr
    - Not optimal for ConvL



Experimental Results – Multi-Resolution

- Q=8 is optimal
  - Consistent with human system
- Multi-resolution helps
  - Best Q=(8,13)
- Time+Frequency scattering helps
  - Not if Q is too low!

Feature	WER Time Scat.	WER Time+Freq. Scat.
log-mel baseline	15.9	15.9
S_1+S_2 (Q=1)	20.5	25.0
S_1+S_2 (Q=4)	16.2	16.3
S_1+S_2 (Q=8)	15.9	15.5
S_1+S_2 (Q=13)	16.1	15.7
S_1+S_2 (Q=1,8)	15.7	15.5
S_1+S_2 (Q=1,13)	15.5	15.5
S_1+S_2 (Q=4,13)	15.6	**15.1**
S_1+S_2 (Q=8,13)	15.3	**15.1**
S_1+S_2 (Q=1,4,13)	15.7	-

Dimensionality Reduction of S_1 & S_2

- Dim. Reduction methods ...
  - $S_2 \rightarrow$ PCA & LDA
  - $S_1 \rightarrow$ Linear bottleneck

Feature	WER	Params
Baseline S_1, tf+S_2, tf (Q=4,13)	15.1	26.5M
S_1, tf+ **pca128**(S_1, f,S_2)	15.2	14.1M
S_1, tf+ **pca256**(S_1, f,S_2)	15.2	15.5M
S_1, tf+ **lda128**(S_1f,S_2)	**15.1**	**14.1M**

- **Conclusion**
  - Identical results with a smaller network

Feature	WER	Params
Baseline S_1, tf+S_2, tf (Q=4,13)	15.1	26.5M
S_1, tf+ lda128(S_1f,S_2)	15.1	14.1M
S_1, tf,**bn=128**+ lda128(S_1f,S_2)	15.4	10.0M
S_1, tf,**bn=256**+ lda128(S_1f,S_2)	**15.1**	**10.8M**

Speaker Adaptation

- VTLN helps!
  - ONLY for S_1 (S_2 unwarped)
- fMLLR & i-vector help!
  - Extra input stream to the FC
    - Do not obey locality
  - More effective than VTLN!
- Using 2xConv Layers help!

Feature	WER no VTLN	WER with VTLN
log-mel	15.9	15.4
S_1+S_2, time+freq, Q=8	15.5	15.0
S_1+S_2, time+freq, Q=4,13	**15.1**	**14.7**

Feature	WER
log-mel +fMLLR+ivectors	13.9
S_1+S_2, time+freq, Q=4,13	**13.4**

Feature	WER
joint CNN/DNN	**13.4**
DNN	14.2

Experimental Results

- Sequence training (after CE) improves the results
- Gain carries over to larger data (50h → 430h)
- Comparing multiQ DSS with log-mel; is it fair?

English Broadcast News, **50h**

Feature	WER
log-mel	12.5
S_1+S_2, time+freq, Q=4,13	**12.0**

English Broadcast News, **430h**

Feature	WER
log-mel	14.2
m1+m2, time+freq, mulitQ	13.2

What are m1 and m2?

“Log-Mel+MFCC” vs DSS

- S_1 and log-mel have identical WER!
- S_2 slightly helps (15.4 → 15.2)
- Frequency scatter slightly helps (15.2 → 15.0)
- Frequency scatter effect is similar to MFCC
- MultiQ “log-mel+MFCCs” **match** DSS with all bells & whistles!

Feature	WER
log-mel, Q=8	15.4
S_1, time scatter, Q=8	15.4
S_1 +S_2 time scatter, Q=8	15.2
$S_1 + S_2$ time+freq scatter, Q=8	15.0
log-mel+mfcc, Q=8	15.0
$S_1 + S_2$ time+freq scatter, Q=4,13	14.7
log-mel + mfcc, Q=4,13	14.6

INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Deep Scattering Power Spectrum Features for Robust Speech Recognition

Neethu M. Joy [1], *Dino Oglic* [1], *Zoran Cvetkovic* [1], *Peter Bell* [2], *and Steve Renals* [2]

[1] Department of Engineering, King's College London, UK
[2] Center for Speech Technology Research, University of Edinburgh, UK
{neethu.joy, dino.oglic, zoran.cvetkovic}@kcl.ac.uk, {peter.bell, s.renals}@ed.ac.uk

INTERSPEECH 2020
OCTOBER 25-29/ SHANGHAI , CHINA
SHANGHAI INTERNATIONAL CONVENTION CENTER

This paper ...

- Investigates usefulness of DSS (S_1 and S_2) for robustness ASR
- Replaces modulus with squared modulus non-linearity
- Comparison with similar architectures

Replace Modulus with Squared Modulus (1)

- Amplifies strong coefficients
  - may improve robustness + better speech/noise separation
- Amplifies *sparsity*

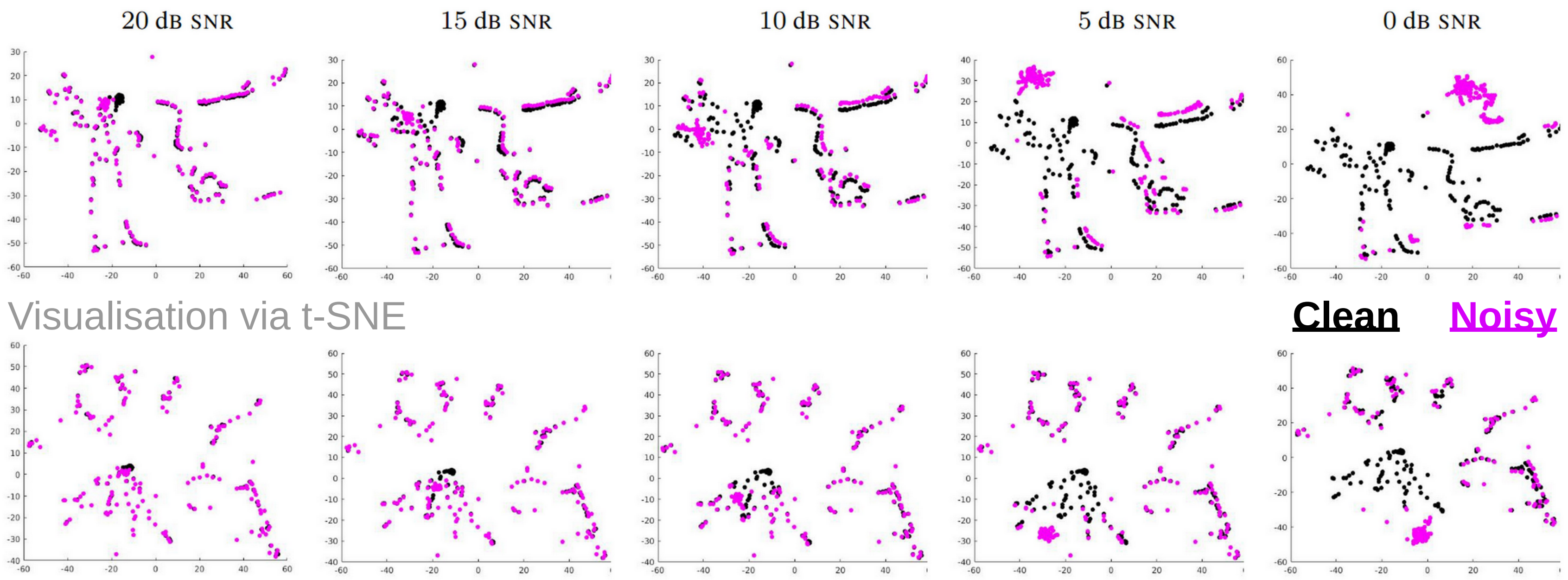
$$\hat{S}_1(t,\lambda_1) = \left|x(t) * \psi_{\lambda_1}(t)\right|^2 * \phi(t)$$

$$\hat{S}_2(t,\lambda_1,\lambda_2) = \left|\,|\,x * \psi_{\lambda_1}(t)\,|^2 * \psi_{\lambda_2}(t)\right|^2 * \phi(t)$$

Replace Modulus with Squared Modulus (2)

Standard DSS
(| . |)

Proposed
(| . |^2)



Visualisation via
t-SNE

| . |^2 is less variant to the additive noise (Guassian) at different SNRs.

Experimental Results – Aurora-4 Clean

Very good WER for this task!

FEATURES	A_1	B_{2-7}	C_8	D_{9-14}	AVG_{1-14}
$DSPS_1$	2.76	13.83	7.74	17.90	14.35
$DSPS_1 + DSPS_2$	2.58	**11.14**	**6.89**	**14.33**	**11.59**
DSS_1 [5]	2.62	14.72	7.89	19.07	15.23
$DSS_1 + DSS_2$ [5]	2.61	11.95	7.33	15.33	12.40
$FBANK_{40}$ [4]	2.65	13.75	7.96	16.89	13.89
$FBANK_{60}$ [4]	2.54	13.06	8.33	17.08	13.69
$FBANK_{80}$ [4]	2.69	12.04	8.03	16.19	12.86
$FBANK_{100}$ [4]	**2.52**	12.60	7.60	16.52	13.20

* Squared modulus → Helps! → 0.9, 0.8% abs

* Second-order features → Helps! → 2.8, 2.8% abs

Experimental Results – Aurora-4 Multi (1)

FEATURES	A_1	B_{2-7}	C_8	D_{9-14}	AVG_{1-14}
$DSPS_1$	2.97	5.88	6.71	15.96	10.05
$DSPS_1 + DSPS_2$	2.73	**5.20**	**4.73**	**14.15**	**8.83**
DSS_1 [5]	2.99	5.69	6.56	15.95	9.96
$DSS_1 + DSS_2$ [5]	2.86	5.45	6.11	15.08	9.44
$FBANK_{40}$ [4]	3.06	6.08	7.10	16.09	10.23
$FBANK_{60}$ [4]	2.90	5.72	6.46	15.65	9.83
$FBANK_{80}$ [4]	2.88	5.58	5.92	15.22	9.55
$FBANK_{100}$ [4]	**2.69**	5.33	5.74	15.26	9.43

* Squared modulus [S_1] → WER → Slight WER increase
* Second-order features → Helps! → 1.2, 0.5% abs

Experimental Results – Aurora-4 Multi (2)

- Multi-Resolution is useful but should not be overdone!
  - Q = {1,4,8,13} is the worst!
  - Best multi-resolution results → Q = {4,13}

- Comparable results with other complicated DNNs

ARCHITECTURE	CNN DEPTH	AVG_{1-14}
$\text{DSPS}_1 + \text{DSPS}_2$ (MULTI-RESOLUTION SCATTERING)		
$Q = \{8\}$	3	8.83
$Q = \{1, 4, 13\}$	3	8.76
$Q = \{1, 4, 8, 13\}$	3	8.94
$Q = \{4, 13\}$	3	**8.64**
FBANK BASELINES		
FMLLR + MLP	-	10.21
VD6CNN [23]	6	10.34
VD10CNN [23]	10	8.81
M-OCT CNN [24]	15	**8.31**

Wrap-up

- Deep scattering spectrum (DSS) is a cascade of wavelet (linear) and modulus (non-linear) transforms
- Advantages: translation invariant, Lipschitz stable & preserves information
- First-order coefficients are similar to filterbank features
- [Novelty] Higher-order aims at recovering lost info in lower level; sparse
  - Usually only first (S_1) and second (S_2) orders are used
- DSS has similar hierarchical structure to CNNs but involves no learning
- Frequency scattering and multi-resolution time scattering are helpful
- Performance on ASR task: comparable to classic features + marginal gain
- Suggestions: learn S_1 via parametric CNNs, use CNN+group for S_2

That's It!

- Thanks for Your Attention!
- Q/A

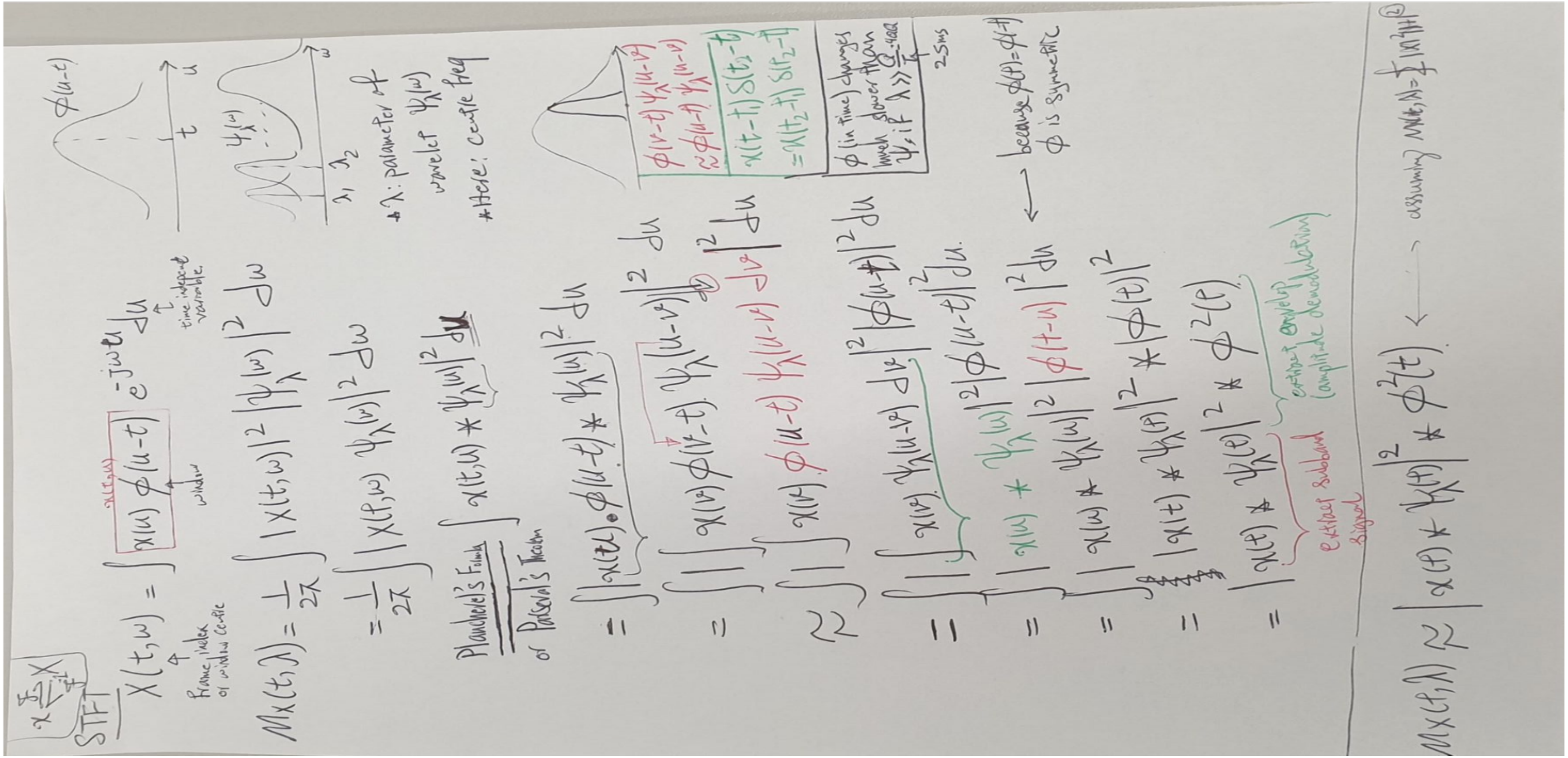
- Appendix A: Proof of

$$Mx(t,\lambda_i) = \int_\omega |X(t,\omega)|^2 |H(\omega;\lambda_i)|^2 \; d\omega$$

- Appendix B: DSS vs ...

$$\approx |x(t) * h(t;\lambda_i)|^2 * \phi^2(t)$$

Appendix A: Proof



Appendix B: DSS vs Modulation Spectrum

Speech Communication 25 (1998) 117-132

Robust Speech Recognition Using the Modulation Spectrogram

Brian Kingsbury, Nelson Morgan and Steven Greenberg