Evaluating Deep Scattering Spectra with Deep Neural Networks on Large Scale Spontaneous Speech Task

Conference Paper · April 2015

DOI: 10.1109/ICASSP.2015.7178832

Petr Fousek

IBM

Evaluating Deep Scattering Spectra with Deep Neural Networks on Large Scale Spontaneous Speech Task

Petr Fousek, Pierre Dognin, Vaibhava Goel

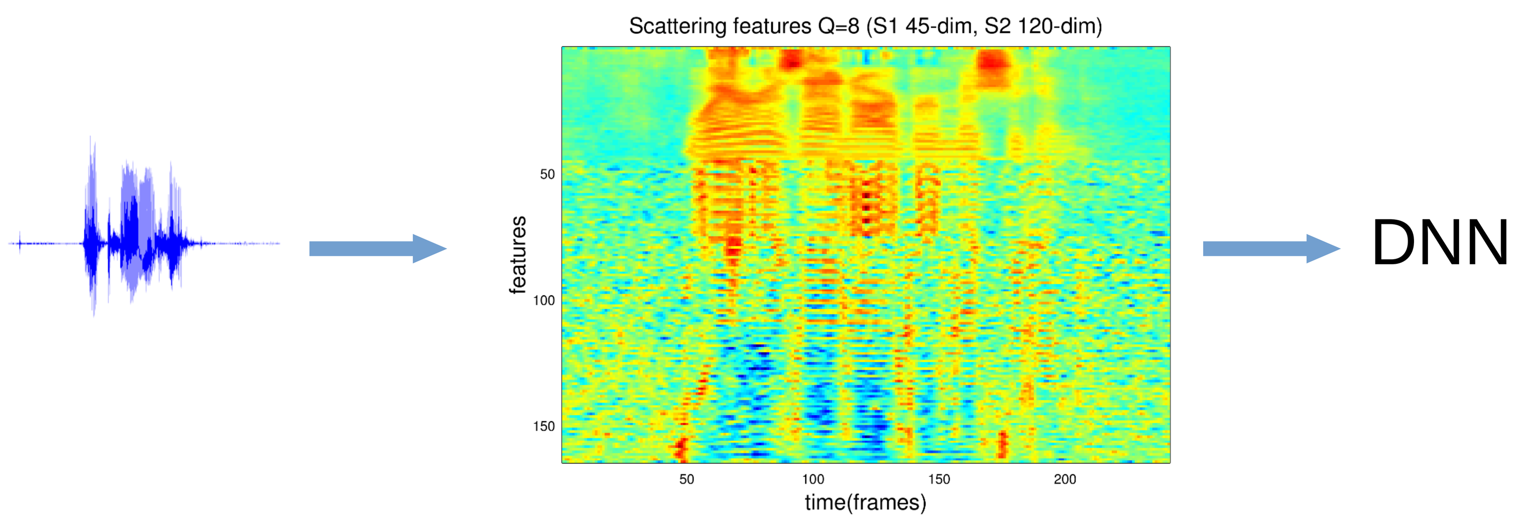
Abstract

- Take a Deep Neural Network **(DNN)** acoustic model, replace common log mel features on its input with with Deep Scatering Spectral features **(DSS)** and look how they compare on spontaneous speech.
- Study how to present DSS features to the DNN and find out which features perform the best.
- To be fair, compare models of the same size.

→ DSS features outperform log mels in all conditions though by a small margin.

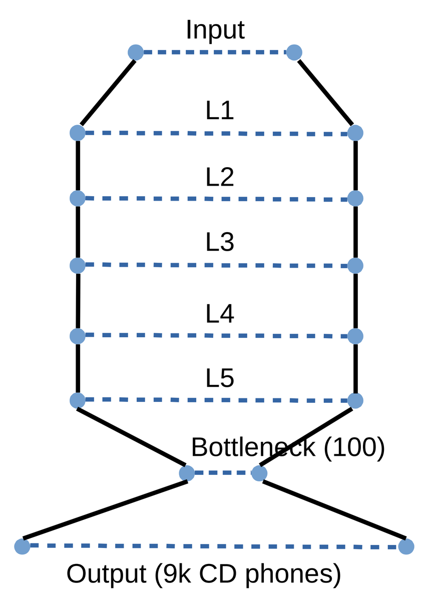
Deep Scatering Spectrum. Why?

- DSS decomposes audio into frequency band-limited signals by wavelet transform.
- Wavelets produce logarithmic frequency output which is what we want (because it works well for log mel or MFCC.).
- No information is lost by binning linear DFT spectra by a filter bank like it is in log mel.
- Wavelet filters for DSS provide features robust to temporal (and frequency) distortions*.
- Depending on the order of the transform, DSS allows for perfect reconstruction so it can encode arbitrary level of detail.



DNN Acoustic Model

- Feed-forward MLP, sigmoids on 5 layers, linear bottleneck, softmax on CD phone targets.
  $[N - 5 * hidN - 100 - 9000]$
- Hybrid setup (DNN gives emission probs. for HMMs).
- Growing layer by layer (x-entropy), then MPE (sequence-level).



training set	feature dim.	hidN	# parameters
125-h	93*11	1024	6.2M
622-h	93*11	2048	20M

Table : Baseline DNN dimensions and model sizes.

Data

- Mobile queries & messages, 3 sec on average, 16 kHz Speex.
- **622 hrs** or **125 hrs** train set, 11 hrs dev set (for DNN), 7 hrs test set.

Features

log mel
31-dim $+\Delta + \Delta^2$
$\pm$**5** frames context = 1023 features

DSS
feature set as described here →
(S_1 *by default go with* Δ, Δ^2)
$\pm$**5** frames context of all

Multi-resolution with Q=1, Q=8

$$S_1 + LDA_{26}, Q8,1$$

First order DSS features

LDA applied on 2[nd] order DSS, retain 26 bases

filter density	# S_1 features	# S_2 features
Q1	10	36
Q4	27	86
Q8	45	120
Q13	63	148

$[S_1 + \mathrm{LDA}_{26}, Q8, 1]$ = 3*(45+10)+26 = 191 features

What DSS Features To Use

- Do we need delta features for DSS? Or equivalent longer context?
- Do second-order DSS (S_2) help?
- Can we compress S_2 with LDA?

features	dim to DNN	PER	WER	hidN*
S_1 (no delta), Q8	45*11	49.6 (49.5)	13.8 (13.7)	1084
S_1 (no delta), Q8	45*15	45.0 (44.9)	13.3 (13.3)	1063
S_1, Q8	135*11	44.7 (44.8)	13.3 (13.1)	975
$S_1 + S_2$, Q8	255*11	44.9 (45.1)	13.0 (13.3)	848
S_1+LDA$_{13}$, Q8	148*11	44.4 (44.5)	13.2 (13.2)	960

* for *same-size* DNNs

Table : Values in parentheses are for *same-size* DNNs. 125-h task, CE training.

Feature normalization

- We **always** apply global mean&var normalization (for DNN).
- **log mel** – per-utterance mean normalization by default (**uttMN**)
- **DSS** – per-utterance L2-normalization on PCM by default (**L2 PCM**)
  PCM scaled by inverse of $\sqrt{\frac{1}{N}\sum_N x_n^2}$

features	norm.	PER	WER
S_1 + LDA$_{13}$, Q8	L2 PCM	44.4 (44.5)	13.2 (13.2)
S_1 + LDA$_{13}$, Q8	raw	44.6 (44.6)	13.3 (13.2)
S_1 + LDA$_{13}$, Q8	uttMN	43.8 (43.9)	13.4 (13.6)
S_1 + LDA$_{13}$, Q8	uttMVN	44.2 (43.9)	13.7 (13.6)
log mel (base)	uttMN	46.2	13.7
log mel	raw	47.3	13.6
log mel	L2 PCM	46.9	13.4

Table : *same-size* DNNs in parentheses have hidN=960. 125-h task, CE training.

Second-order features are more sensitive to L2-norm. S_1 + S_2, Q8:

norm.	WER
L2 PCM	13.0 (13.3)
raw	13.4 (13.5)

*hidN = 848

Acknowledgment: We thank Steven Rennie and Tara Sainath for valuable insights.

Filter bank resolution

- Previous study on S_1 features [Sainath et al. '14] says Q=8 (45-dim) is enough.
- Does $S_1 + \mathrm{LDA}$ show a different trend?

features	dim	PER	WER
S_1+LDA$_5$, Q1	35	51.1	16.5
S_1+LDA$_9$, Q4	90	44.5	13.2
S_1+LDA$_{13}$, Q8	148	44.4 (44.5)	13.2 (13.2)
S_1+LDA$_{21}$, Q13	210	44.8 (44.9)	13.3 (13.3)

... **log mel** behaves similarly, best performance between Q=4 and Q=8.

Multi-resolution filter banks

- Multi-resolution for DSS makes sense (sparser temporal resolution $\Longleftrightarrow$ finer frq. resolution) → complementarity.
- Concatenate S_1 features, apply LDA on merged S_2 features.
- **log mel** has limited complementarity (same DFT input).

fea	dim	PER	WER	hidN*
log mel (uttMN)	93	46.2	13.7	1024
L2-norm on audio				
log mel	93	46.9	13.4	1024
log mel, Q8,4,1	246	46.8 (47.1)	13.3 (13.4)	857
S_1+LDA$_5$, Q1	35	51.1	16.5	
S_1+LDA$_9$, Q4	90	44.5	13.2	
S_1+LDA$_{13}$, Q8	148	44.4 (44.5)	13.2 (13.2)	960
S_1+LDA$_{11}$, Q4,1	122	44.5 (44.4)	13.1 (13.0)	990
S_1+LDA$_{15}$, Q8,1	180	44.1 (44.3)	12.8 (13.0)	925
S_1+LDA$_{28}$, Q8,4	244	44.0 (44.2)	12.6 (13.0)	859
S_1+LDA$_{26}$, Q8,4,1	272	43.8 (44.1)	12.7 (12.9)	832
No normalization				
log mel	93	47.3	13.6	1024
log mel, Q8,4,1	246	47.2 (47.5)	13.6 (13.7)	857
S_1+LDA$_5$, Q1	35	51.5	16.8	
S_1+LDA$_9$, Q4	90	44.7	13.3	
S_1+LDA$_{13}$, Q8	148	44.6 (44.6)	13.3 (13.2)	960
S_1+LDA$_{11}$, Q4,1	122	44.6 (44.6)	13.4 (13.3)	990
S_1+LDA$_{15}$, Q8,1	180	44.5 (44.2)	13.1 (13.1)	925
S_1+LDA$_{28}$, Q8,4	244	44.3 (44.2)	13.0 (13.1)	859
S_1+LDA$_{26}$, Q8,4,1	272	44.2 (44.3)	13.1 (13.1)	832

* for *same-size* DNNs

Results on Full Data Set

- **log-mel** baseline carefully tuned, hidN=2048, 20M parameters.
- On cross-entropy, **DSS** is better by 4% relative than log mel, on sequence-training by 3% relative.

fea	dim	PER	WER	ST WER
log mel (uttMN)	93	39.3	11.5	10.0
L2-norm on audio				
S_1+LDA$_{26}$, Q8,4,1	272	39.5 (38.8)	11.2 (11.1)	9.7 (9.7)
S_1+LDA$_{15}$, Q8,1	180	39.1 (39.3)	11.0 (11.2)	9.8 (9.8)
No normalization				
S_1+LDA$_{26}$, Q8,4,1	272	39.5 (39.2)	11.3 (11.3)	10.0 (9.9)

Table : 622-h data set. WER for Cross-Entropy and Sequence-Training.

IBM Prague Speech Lab (CZ), IBM T. J. Watson Center (USA)
petr.fousek@cz.ibm.com, {pdognin,vgoel}@us.ibm.com