On Spectral and Temporal Feature Encoding Behaviour in Stacked Architectures

Vaibhav Singh
NYU, USA
vaibhav.singh@nyu.edu

Vinayak Abrol
IIIT Delhi, India
abrol@iiit.ac.in

Karan Nathwani
IIT Jammu, India
karan.nathwani@iitjammu.ac.in *

Abstract

Acoustic models typically employ production and perception based short-term features. In the context of deep models the acoustic information is hierarchically combined either 1) across frequency bands followed by temporal modelling similar to cepstrum features; or 2) across temporal trajectories followed by combination across spectral bands similar to relative spectra (RASTA) features. Such a processing pipeline is often implemented using low-rank methods to achieve low-footprint compared to SOTA models involving simultaneous spectral-temporal processing. However, very few attempts have been made to address the question of if and how such deep acoustic models flexibly integrate information from spectral or temporal features. In this work with the help of an Large vocabulary continuous speech recognition (LVCSR) case study, the geometry of loss landscape is used as a visualisation tool to understand the link between generalization error and spectral or temporal feature integration in learning task-specific information.

1 Introduction

Acoustic models have undergone a rapid evolution; from mixture model and HMM based Meyer and Schramm [2006] to hybrid DNN-HMM Golik et al. [2015] and more recent end-to-end models Gehring et al. [2013], Rao et al. [2017] achieving state-of-the-art (SOTA) performance in various tasks. For any deep model, one of the most important part is its front-end feature engineering pipeline, which uses some kind of spectral or temporal processing. Typically, time-frequency (TF) representations (e.g., spectrograms) extracted from raw audios are fed to the neural networks (NN) which embeds them to higher dimensional spaces, from which inferences can be drawn be it classification, language prediction or speech recognition. Two of the most popular front-end feature processing pipeline used in deep learning paradigm are *Regular Convolutions* and *Compressed Convolutions*. There are many types of regular convolutions which vary depending on the implementation requirement, like 1D, 2D, 3D, with others variants such as deformed, dilated, point-wise, and grouped convolutions. Compressed convolutions includes Depth-wise Separable Convolutions (DSC) Chollet [2017] and Low Rank Convolutions (LRC) Tai et al. [2016], Abrol et al. [2019]. In DSC temporal trajectories are processed followed by combination across spectral bands. In contrast, spectral trajectories are processed first in LRC followed by combination across temporal dimension. These methods have an added advantage of reducing the number of parameters, making the model robust and fast to train. While different studies claim different performance gains in different training and architectural setups, a very few attempts have been made to cater to questions like *Which feature pipelines really capture the generalised relationship between different spectral and temporal features?* or *When a particular feature processing method should be favoured over the other?*. The performance gap is often bridged by either feeding in more data or upgrading the model to multi-modal setting; both of

*This work is supported by SERB-Startup Reseach Grant (SRG/2021/002348), Govt. of India and Infosys Centre for AI, IIIT Delhi.

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

which makes it even harder to understand the modelling capability and behaviour of the underlying acoustic model. While there exists some recent explainbility methods for acoustic models, they are mainly restricted to analysis of first layer in raw-waveform models Gupta and Abrol [2022] or using example specific gradient salience maps that are often redundant and can't explain the generalization behaviour of the whole model Muckenhirn et al. [2019].

In this work we aim to understand how spectral or temporal feature integration helps in recognising task-specific information. With a case study on LVCSR (Large vocabulary continuous speech recognition), we compared the performance of both DSC and LRC based front-end encoder in an end-to-end system. To make our work more exhaustive both types of inputs raw-waveform and spectrograms along with deep and shallow architectures are considered. The generalization behaviour of these models is studied and explained using network's *Loss Landscape* Li et al. [2018]: which helps in visualising the geometry of loss surface irrespective of the architecture. The main takeaways from our study are

1. Moving from shallow networks to deep networks, the superiority of one feature pipeline over the other vanishes for large scale task and the generalisation error improves. But its somewhat comparable for both spectral and raw-waveform inputs.
2. Deep models have flat minimizers and are less chaotic as compared to sharper one with chaotic behaviour in case of shallow models. Further, the sharpness of loss landscape correlates well with the model's performance in practice.

All the experiments are reproducible & implementation is available on GitHub[2]. We encourage readers to see link for animated plots towards better understanding and readability.

2 Network Visualisation using Loss Landscape

Training a NN requires optimizing a high-dimensional non-convex loss function whose shape is greatly affected by choice of architecture, initializer, regularizer, optimizer etc Goodfellow et al. [2015], Garipov et al. [2018]. Since, loss functions live in a high-dimensional space, most existing methods perform visualization of loss landscape in low-dimensions. In 2D case, one chooses two direction vectors θ_1 and θ_2 that act as perturbation to the original network weight θ^* vector to plot a function of the form

$$f(\eta, \lambda) = L(\theta^* + \theta_1\eta + \theta_2\lambda), \tag{1}$$

where η and $\lambda \in$ [-1,1]. In this work, the choice of θ_1 and θ_2 follow the filter-wise normalisation technique as proposed in Li et al. [2018]. Inferences such as impact of network depth on sharpness of minimizer can aid in explaining the performance, learning behaviour and generalization of the model.

3 Experimental Setup and Results

Our experiments are performed on the Librispeech Panayotov et al. [2015] dataset which consists of 970 hours of labelled speech with end-to-end framework based on ContextNet Han et al. [2020]. This study relies on the TensorflowASR implementation where we modify the encoder pipeline, keeping the remaining NN architecture and training strategy such as optimizer, learning-rate scheduler and augmentation unaltered. To understand the encoder learning behaviour exclusively an external language model is not employed and all experiments are performed using a single V100 GPU.

3.1 Model

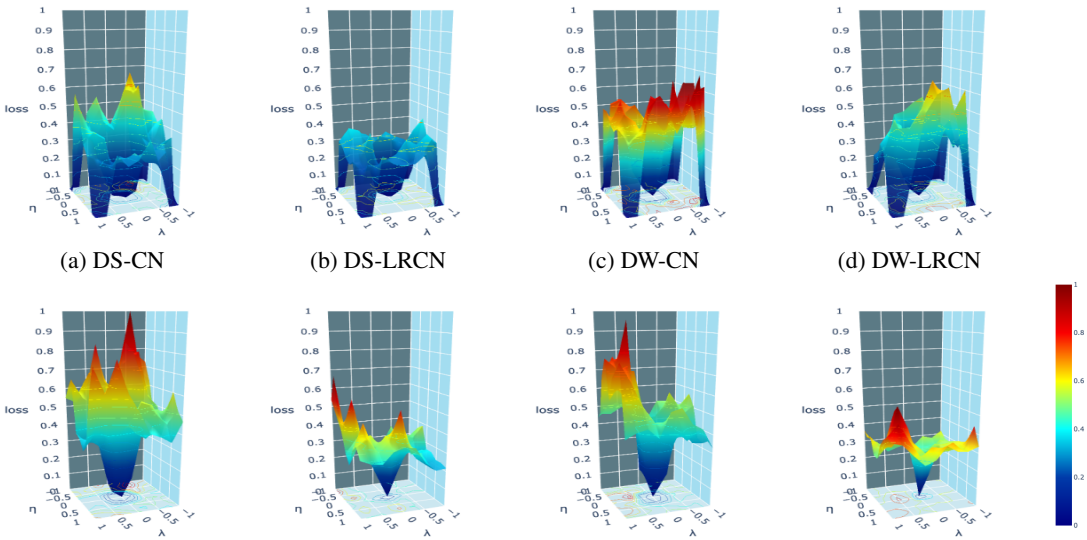
In this work, we consider the end-to-end framework based on ContextNet Han et al. [2020], which consists of three components: front-end convolutional neural network (CNN) feature encoder, long-short term memory (LSTM) based label encoder, and a recurrent neural network-transducer (RNN-T) based joint decoding network to combine the former two. Our primary interest is in the CNN encoder, which is trained on log-mel spectral features or raw-waveforms as inputs. For raw waveform models, we augment the encoder with a 1D CNN layer, where parameters (kernel size-400; stride-160 and #filers-80) are chosen such that the dimensions of output feature maps are equivalent to

[2] `https://github.com/Cross-Caps/STFADE`

the 80 dimensional log-mel spectrograms. Due to computational constraints, only reverberation & additive noise augmentation is performed when training with wave inputs, and we expect the overall performance to improve with more augmented data. Broadly we consider two types of encoder design based on the the types of convolutions they employ. 1) vanilla ContextNet (CN) with DSC, and 2) Low-Rank ContextNet (LRCN) with LRC Abrol et al. [2019]. Furthermore, we present a comparison of deep and shallow architecture, with former having 23 convolutional blocks stacked with skip connections. Each block has 5 convolution layers as described in Han et al. [2020]. To obtain a shallow configuration, these blocks are replaced with single convolutional layers without any skip connections.

Table 1: Comparison of WER for different models averaged over 10 trials.

Input	NN Type	Model	#Param (M)	WER (test-clean/other)
Spectral	*Deep*	CN	10.82	2.6 / 7.0
		LRCN	10.85	2.49 / 7.1
	Shallow	CN	6.29	9.4 / 12.2
		LRCN	6.30	9.7 / 13.1
Wave	*Deep*	CN	10.85	2.7 / 7.1
		LRCN	10.88	2.4 / 7.0
	Shallow	CN	6.32	9.5 / 12.3
		LRCN	6.33	9.7 / 13.2



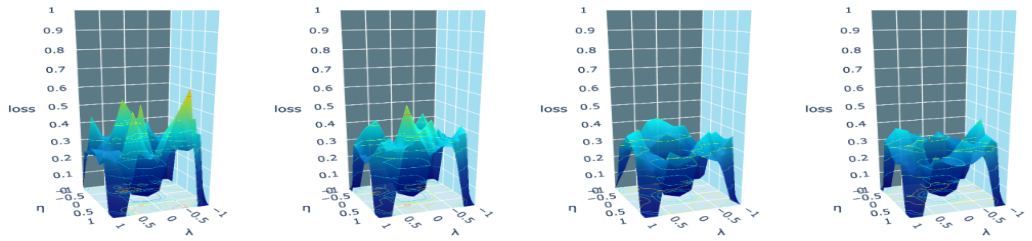
(a) DS-CN (b) DS-LRCN (c) DW-CN (d) DW-LRCN

(f) SS-CN (g) SS-LRCN (h) SW-CN (i) SW-LRCN

Figure 1: Loss landscapes on testset for various trained acoustic models. 'DS/DW' and 'SS/SW' denotes deep or shallow models with spectrogram or raw-waveform as inputs

3.2 ASR Results on LibriSpeech

Table1 reports different acoustic models that are studied in this work, their learnable parameters and corresponding word error rates (WER) on test-clean/other data splits averaged over 10 trials. It is observed that in case of deep spectral models both CN and LRCN models achieve comparable but lowest WER. In contrast shallow models struggle with their performance as compared to the deep encoders, highlighting their limited capacity to generalize well. These results are in line with SOTA performance reported in earlier studies leading to strong baselines for our study. Further, models with wave inputs achieve comparable WER, and we expect the performance would match to that of their spectral counterparts in presence of a better augmentation strategy. These results demonstrates that performance gains with temporal or spectral first processing based encoders vanishes as the depth



(a) Epoch 3 (b) Epoch 11 (c) Epoch 14 (d) Epoch 20

Figure 2: Evolution of loss landscape of DS-LRCN model during training. Best viewed in colour.

of the encoder increases. Apart from the basic CNN structure there are other factors such as kernel size/stride, normalization, weight regularization, dropout, skip-connections and loss functions which may impact the overall performance. An extensive study comparing various approaches with respect to these choices is out of the scope of this paper. In the next section we will investigate the obtained WER using the loss landscapes of these models to explain and bring out some meaningful insights.

3.3 Loss Landscapes of Different models

Figure1 shows the loss landscapes for different models considered in Table1, where only the parameters of the encoder are modified keeping all other parameters of the model fixed. Due to space constraints, only visualization for final trained model are plotted. Consistent with earlier studies Li et al. [2018], Chaudhari et al. [2017], we observed that the sharpness of loss landscape minimum correlates well with the respective WER reported in Section 3.2. Deep models with stacked blocks having multiple convolutional layers per block have flat minimizers and are less chaotic as compared to sharper minimizer with chaotic behaviour in case of shallow models. These results coincides with the model's training ability in practice and ultimately to observed generalization error in terms of WER on testsets. It can also be infered that overall LRCN both for deep and shallow models, attains a lower loss than CN model, which shows that models with LRC front-end are more robust to weight perturbations/updates between successive gradient updates. The evolution of loss landscape in Fig. 2 for LRCN model demonstrate this behaviour where the landscape transitions from more to less chaotic over training. Beyond just performance, these observations opens up future avenues such as to understand robustness against adversarial examples and the role of temporal context in learning invariant acoustic representations. Similarly, the feature redundancy makes room for further network optimisation such as in terms of model compression and optimal depth. Finally, among shallow models, we observe very sharp minimizer for LRCN compared to CN model leading to results in contradiction to earlier studies that shown LRC consistently outperforms DSC on a variety of datasets Abrol et al. [2019]. We argue that this behaviour is due to very different interaction of the front-end CNN encoder with the transducer and/or RNN-decoder in a end-to-end system compared to isolated HMM-decoder based system, and this aspect demands a further analysis in future work.

4 Discussion and Conclusion

In this paper we made an attempt to understand feature integration and modelling in acoustic models with different front-end processing pipelines. Through a case study on LVCSR task various modelling choices are discussed and compared. In particular, we showed that for deep networks the superiority of one feature pipeline over the other vanishes for a large scale task. We also demonstrated that the bottleneck in the performance gap between models trained on spectrograms and raw waveforms is due to sub-optimal choice of first layer parameters and lack of augmentation. Using loss landscape visualizations we showed that the sharpness of minimizer correlates well with the respective generalization error of a model, however further analysis is required to understand the impact of interactions between feature encoder and decoder on performance of a system. Our study has implications towards building an efficient model for edge applications for a given speech/audio task, where the goal is to select an optimal spectral, temporal or specto-temporal acoustic modelling pipeline which will result in the best accuracy even on out-of-domain data.

References

Vinayak Abrol, S. Pavankumar Dubagunta, and Mathew Magimai.-Doss. Understanding raw waveform based CNN through low-rank spectro-temporal decoupling. Technical report, Idiap, October 2019.

P. Chaudhari, Anna Choromanska, S. Soatto, Yann LeCun, C. Baldassi, C. Borgs, J. Chayes, Levent Sagun, and R. Zecchina. Entropy-SGD: Biasing gradient descent into wide valleys. In *International Conference on Learning Representations (ICLR)*, pages 1–16, 2017.

François Chollet. Xception: Deep learning with depthwise separable convolutions. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1800–1807, 2017. doi: 10.1109/CVPR.2017.195.

Timur Garipov, Pavel Izmailov, Dmitrii Podoprikhin, Dmitry Vetrov, and Andrew Gordon Wilson. Loss surfaces, mode connectivity, and fast ensembling of dnns. In *International Conference on Neural Information Processing Systems (NeurIPS)*, page 8803–8812, 2018. doi: 10.5555/3327546.3327556.

Jonas Gehring, Yajie Miao, Florian Metze, and Alex Waibel. Extracting deep bottleneck features using stacked auto-encoders. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3377–3381, 2013. doi: 10.1109/ICASSP.2013.6638284.

P. Golik, Z. Tüske, R. Schlüter, and H. Ney. Convolutional neural networks for acoustic modeling of raw time signal in LVCSR. In *INTERSPEECH*, pages 26–30, 2015.

Ian J Goodfellow, Oriol Vinyals, and Andrew M Saxe. Qualitatively characterizing neural network optimization problems. In *International Conference on Learning Representations (ICLR)*, pages 1–20, 2015.

Devansh Gupta and Vinayak Abrol. Time-frequency and geometric analysis of task-dependent learning in raw waveform based acoustic models. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4323–4327, 2022. doi: 10.1109/ICASSP43922.2022.9746577.

Wei Han, Zhengdong Zhang, Yu Zhang, Jiahui Yu, Chung-Cheng Chiu, James Qin, Anmol Gulati, Ruoming Pang, and Yonghui Wu. Contextnet: Improving convolutional neural networks for automatic speech recognition with global context. In *INTERSPEECH*, pages 3610–3614, 2020. doi: 10.21437/Interspeech.2020-2059.

Hao Li, Zheng Xu, Gavin Taylor, Christoph Studer, and Tom Goldstein. Visualizing the loss landscape of neural nets. In *International Conference on Neural Information Processing Systems (NeurIPS)*, page 6391–6401, 2018.

Carsten Meyer and Hauke Schramm. Boosting HMM acoustic models in large vocabulary speech recognition. *Speech Communication*, 48(5):532–548, 2006. ISSN 0167-6393. doi: https://DOI.org/10.1016/j.specom.2005.09.009.

Hannah Muckenhirn, Vinayak Abrol, Mathew Magimai.-Doss, and Sébastien Marcel. Understanding and visualizing raw waveform-based CNNs. In *INTERSPEECH*, pages 2345–2349, September 2019. doi: 10.21437/Interspeech.2019-2341.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and S. Khudanpur. Librispeech: An ASR corpus based on public domain audio books. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210, 2015.

Kanishka Rao, Haşim Sak, and Rohit Prabhavalkar. Exploring architectures, data and units for streaming end-to-end speech recognition with RNN-transducer. In *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 193–199, 2017. doi: 10.1109/ASRU.2017.8268935.

Cheng Tai, Tong Xiao, Yi Zhang, Xiaogang Wang, and E. Weinan. Convolutional neural networks with low-rank regularization. In *International Conference on Learning Representations (ICLR)*, pages 1–11, 2016.