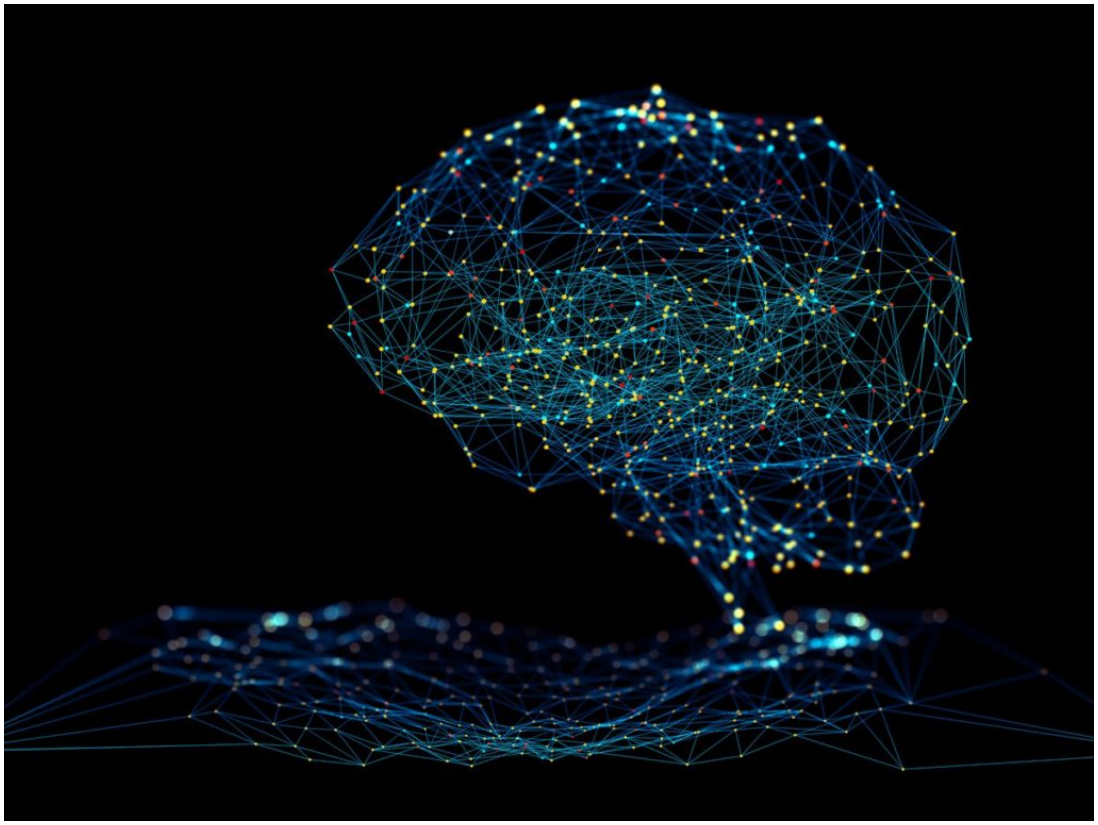
NUANCE | WHAT'S NEXT BLOG

Enterprise R&D, Innovation & Research

# Delivering non-intrusive signal intelligence with deep learning

Innovation at Nuance | Employee Guest Blogger

January 13, 2021



Teaching machines to perceive the overall quality of speech and estimate the underlying acoustic characteristics of the environment in which the signal was captured without a clean reference signal (i.e. non intrusively) is a challenging task that has many real-world applications. These include audio forensics, hearing aids, speaker diarization and automatic speech recognition (ASR). In the context of ASR, an important application is for the purpose of data augmentation and selection, where an algorithm like Non-Intrusive Speech Analysis (NISA) can be used for analyzing field data to estimate the distributions of various acoustic parameters present and then to use those distributions to select appropriate data for ASR model training as well as for the purpose of configuring and automating data augmentation strategies.

**About the author:** Dushyant Sharma is a Principal Research Scientist, working as part of the Central Research and Development organization at Nuance on front end signal processing and acoustic data augmentation. His research interests include non-intrusive speech signal analysis, single and multi-channel speech enhancement and acoustic data augmentation and simulation. Dushyant joined Nuance in 2012 after completing his Ph.D.

in speech signal processing from the Centre for Law Enforcement Audio Research (CLEAR) at Imperial College in London, UK.

---

When a signal is acquired in a room with reflective surfaces and noise sources, the acquired signal can be modeled by convolution with the room impulse response (RIR) plus an additive noise component. The level of reverberation can be characterized by an RIR, from which the Clarity Index (C50) metric can be obtained which has been shown to be well correlated with ASR performance. The level of additive noise can be modeled by the Signal-to-Noise Ratio (SNR). The combined effects of noise and reverberation can be modeled by the Perceptual Evaluation of Speech Quality (PESQ) algorithm, which is an intrusive method requiring the clean reference signal. Another important aspect of the signal is estimating the segments where speech is present, a task commonly referred to as voice activity detection (VAD).

In our paper "Non-Intrusive Estimation of Speech Signal Parameters using a Frame-based Machine Learning Approach", to be presented at EUSIPCO 2020, we propose a multi-task machine learning framework called NISA for non-intrusive acoustic parameter estimation that includes VAD, C50, PESQ and segmental SNR. While most recent methods for non-intrusive speech quality and reverberation parameter estimation operate on a large temporal window size (more than a second long) or at an utterance level, our framework is able to reliably estimate a number of parameters in short windows of length 300ms. The ability to estimate these parameters in short-time windows enables the use of these parameters as additional features in speaker diarization, signal quality assurance and for signal selection in multiple microphone ASR.

In this paper we explore three different feature extraction front-ends (MMF, MFB and PASE) and two different deep learning architectures (CNN and LSTM) for the joint estimation of the acoustic parameters.

The three feature extraction front ends we explore are as follows:

1. MMF – this is a combination of MFCCs and Modulation domain features as described in [1]
2. MFB – this refers to the use of Mel Filter Bank features (similar to MFCCs in [1] but without the DCT). Spectral Augmentation (SA) [3] is further applied (randomly drop 3% of the Mel Channels for a given block of frames).
3. PASE – a Problem-Agnostic Speech Representation (PASE) [2] feature set that is extracted from the waveform of speech using a self-supervised encoder discriminator architecture with several target encoders (including MFCC features). Here we use the pre-trained model provided by the authors.

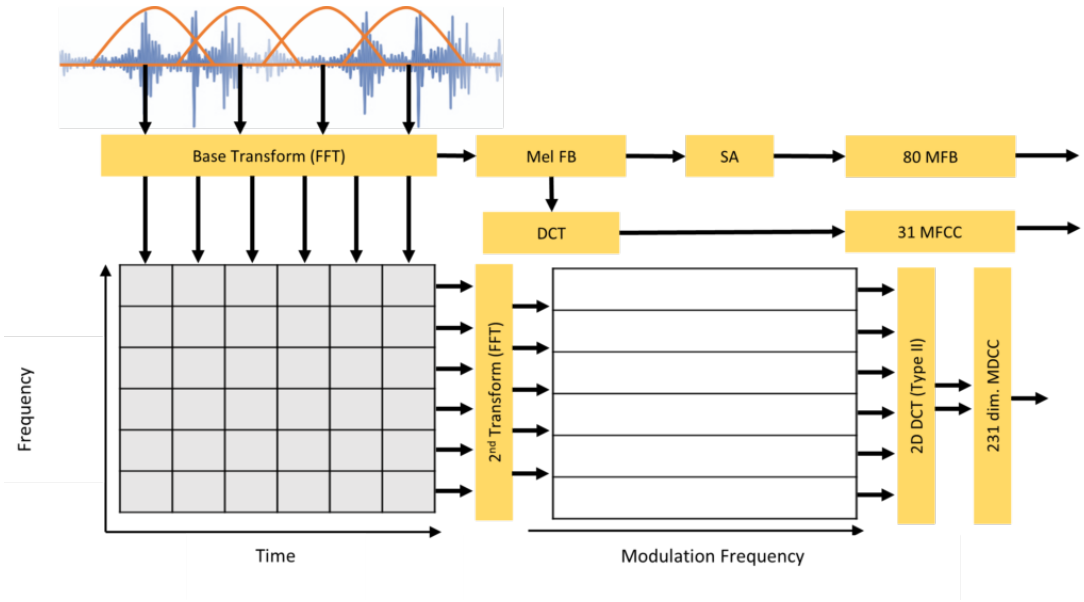



Figure 1 MMF and MFB feature extraction steps
We explore two deep learning frameworks, trained with an RMS cost function and the Adam optimizer [4]

1. LSTM – a recurrent structure [5], which has been shown to be a powerful architecture for modelling time varying features. Our model is composed of three hidden layers, arranged in a 108x54x27 cell topology (for each time step).
2. CNN – a compact Convolutional Neural Network model based on the SwishNet [6] architecture with important modifications:
    - Front end feature MFCC -> MFB + SA
    - Experimented with different filter sizes
    - Added a dropout layer in the architecture

- Output layer modified for regression: linear layer with four output nodes

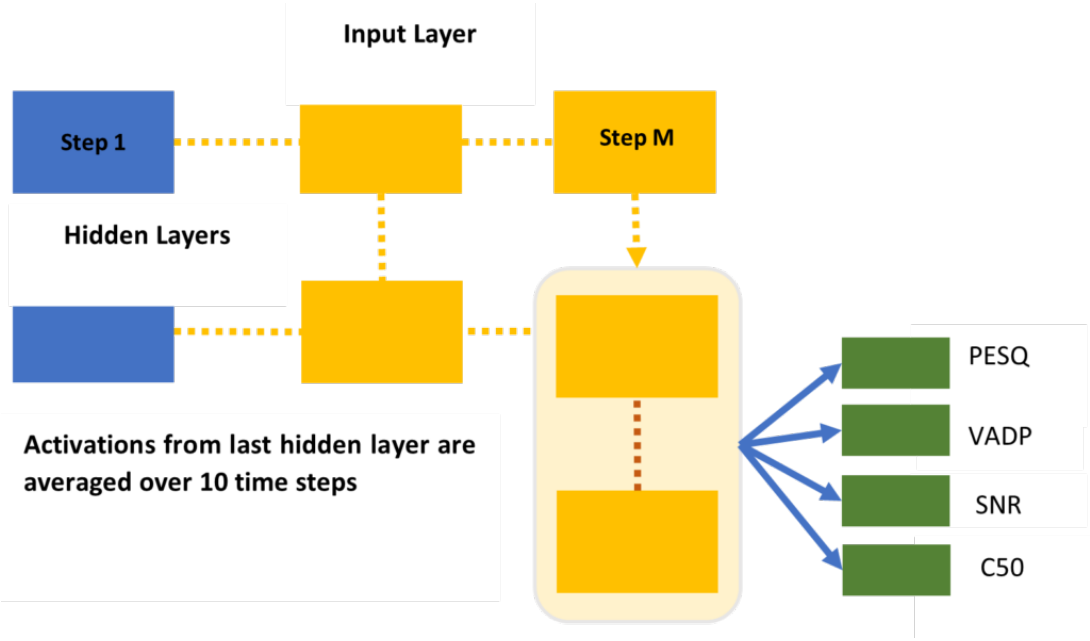



Figure 2 The LSTM architecture.

The different combinations of feature front ends and machine learning architectures provide tradeoffs in performance in terms of computational complexity and estimation accuracy. The systems are trained on simulated data with clean speech from the training partition of the Wall Street corpus, convolved with simulated RIRs and additive noise (ambient, babble, domestic and white). The RIRs are sampled with C50 in the range [0 to 30 dB]. We use two simulated test sets for evaluating the performance of the proposed framework with base speech material from two different sources to the training data (to avoid any overlap in words, speaker or recording system) and the RIRs and noise sources are also separate with no overlap with training data.

| | | | Mean Absolute Error | F1 Score | | | |
|---|---|---|---|---|---|---|---|
| Feature Type | Spectral Augmentation | Model Type | C50 | SNR | PESQ | VAD | Trainable Parameters |
| MMF | N/A | LSTM | 2.9 | 1.8 | 0.3 | 0.92 | 126k |
| MFB | N | LSTM | 3.1 | 2.2 | 0.3 | 0.93 | 126k |
| MFB | Y | LSTM | 3.3 | 2.3 | 0.3 | 0.93 | 126k |
| PASE | N/A | LSTM | 2.8 | 1.4 | 0.4 | 0.93 | 5.9M |
| MFB | Y | CNN | 3.5 | 2.6 | 0.4 | 0.93 | 16K |

Figure 3. NISA parameter estimation results and model complexity.

We show that a low complexity MFB-based feature extraction with spectral augmentation and an LSTM model achieves good performance for acoustic parameter estimation and provides a good trade-off in performance and complexity. We show the performance of this system is stable across different noise types and how different ground truth and estimated parameters correlate with ASR performance (in terms of WER as well as the sub-components such as insertions, deletions etc.). We show, firstly that of the three reverberation level estimation parameters, C50 is the most correlated with WER, and in addition, SNR and PESQ are also highly correlated with WER. Lastly, we show that the C50, SNR and PESQ estimated by the NISA model are also highly correlated with ASR performance.

*Lucia Berger, Carl Quillen and Patrick A. Naylor contributed to the paper and this blog post.*

# References

[1] D. Sharma et. al. “Non-intrusive polqa estimation of speech quality using recurrent neural networks,” EUSIPCO, 2019

[2] S. Pascual et. al. “Learning problem-agnostic speech representations from multiple self supervised tasks,” INTERSPEECH, 2019.

[3] D. S. Park et. al., “SpecAugment: A simple data augmentation method for automatic speech recognition,” INTERSPEECH, 2019.

[4] D. P. Kingma et. al. “Adam: A Method for Stochastic Optimization,” vol. abs/1412.6980, 2014. [Online]. Available: http://arxiv.org/abs/1412.6980

[5] S. Hochreiter et. al. “Long Short-Term Memory,” Neural Computation, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.

[6] M. Hussain et. al. “Swishnet: A fast convolutional neural network for speech, music and noise classification and segmentation,” 2018, [Online]. Available: https://arxiv.org/abs/1812.00149

**Tags:** Research & insights