NON-INTRUSIVE POLQA ESTIMATION OF SPEECH QUALITY USING RECURRENT NEURAL NETWORKS

DUSHYANT SHARMA[1], **AIDAN O. T. HOGG**[2], YU WANG[3], AMR NOUR-ELDIN[1] AND PATRICK A. NAYLOR[2]
[1] NUANCE COMMUNICATIONS INC., USA [2] IMPERIAL COLLEGE LONDON, UK [3] UNIVERSITY OF CAMBRIDGE, UK

Overview

What is speech quality?

Answers the question "How does it sound?"

Subjective methods : panel of listeners judge quality
- Pros: Accurate
- Cons: Expensive & time consuming

Objective methods : estimate using signal processing
- Pros: Cheaper & faster (online processing possible)
- Cons: Less accurate

Motivation

How do objective methods work?

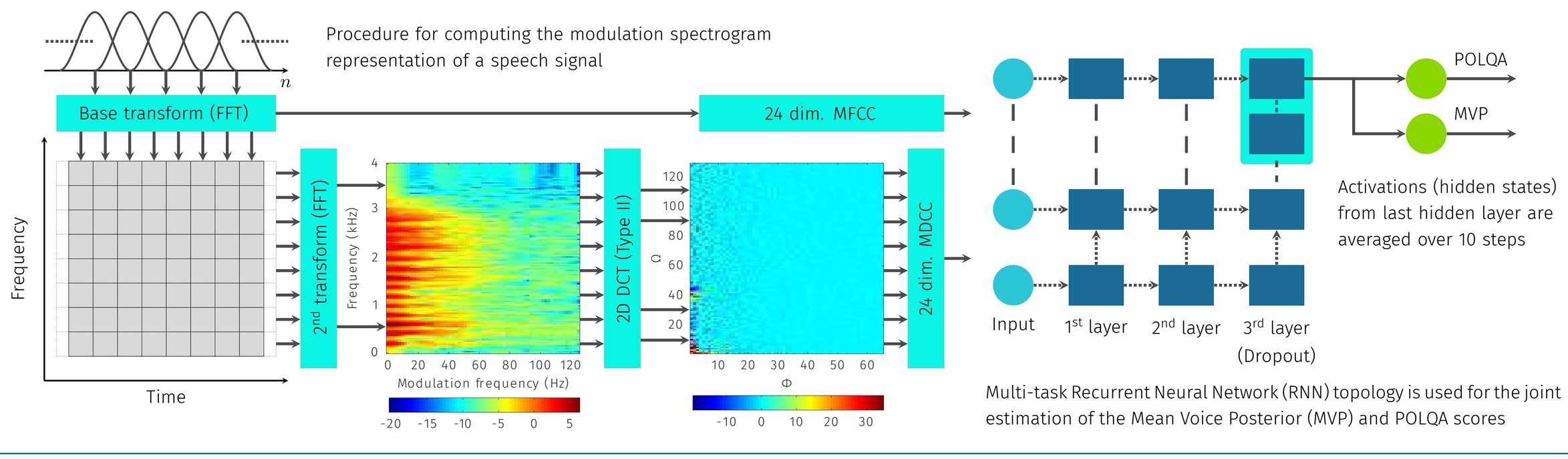
- Intrusive methods: Require both the clean (reference) & degraded signal, e.g., POLQA
- Non-Intrusive methods: Require only the degraded signal

Is there a way to estimate POLQA non-intrusively?

We propose the first short-time, Non-Intrusive Speech Quality Assessment (NISA) estimator that uses a Recurrent Neural Network (RNN) to estimate POLQA scores.

- Use intrusive POLQA algorithm to score artificially corrupted data
- Use RNN to estimate POLQA

Proposed System: Non-Intrusive Speech Quality Assessment (NISA)



Recurrent Neural Network (RNN) Model

Input layer
- Size - 48×1

Hidden layers
- 3 layers of LSTM cells in a $40 \times 21 \times 16$ neuron topology (for each time step)
- In the last hidden layer, the activations are averaged over a window of 10 frames

Output layer
- Two nodes - one node to estimate the POLQA score and a second node to estimate the Mean Voice Posterior (MVP)

Dropout layer
- To avoid over-fitting the model parameters, dropout is applied before the last hidden layer of the RNN

Data

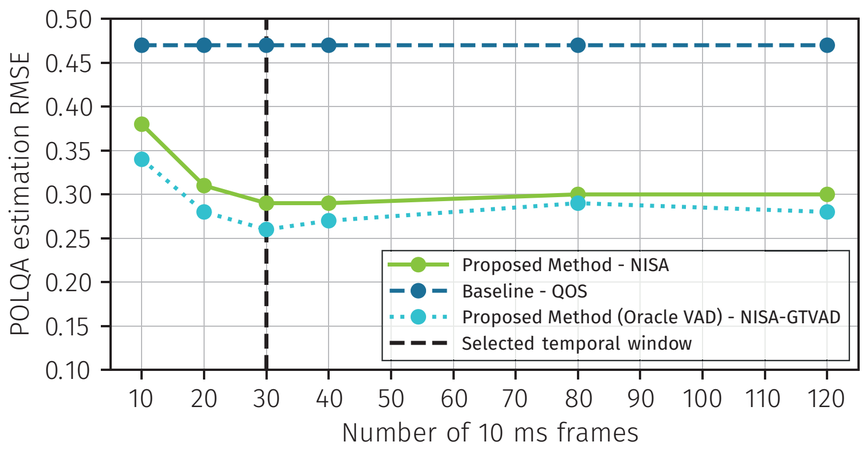
Training data (100 hrs. POLQA labelled)

- Speech: 100hr partition of clean speech from Librispeech corpus
- Room Impulse Response (RIR): 128 RIRs - T60 in [0.1 to 1.25s], C50 in [0 to 30 dB]
- Added noise: SNR range [0 to 30 dB] - Ambient, Babble, Household, Street & Vehicle
- CODEC: 10 conditions - G.711 (a-law), GSM-FR, G.729 (A/B), GSM-AMR (4.75, 6.7, 7.4 & 12.2 kbps) & linear PCM

Test data (15 hrs. POLQA labelled)

- Procedure: Same as above but with unseen speech, RIR & noise data (no overlap)
- RIR & Noise: 32 RIRs & 36 noise sources
- Speech: Test-clean partition from Librispeech & ITU-T P.23 (English, French & Japanese)

Results



POLQA estimation RMSE (averaged over all test sets) using the proposed method (middle line), the baseline QOS [1] method (top line) and the proposed method with oracle VAD used for testing (bottom line)

[1] D. Sharma *et al*, "A non-intrusive PESQ measure," in *2014 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*

MFCC only	MAD	RMSE	R	VAD F1
Libre - English	0.35	0.54	0.70	0.86
P23 - English	0.28	0.43	0.67	0.91
P23 - French	0.22	0.31	0.69	0.89
P23 - Japanese	0.24	0.38	0.70	0.89
Mean	0.27	0.42	0.69	0.89

MFCC + MDCC	MAD	RMSE	R	VAD F1
Libre - English	0.26	0.37	0.86	0.89
P23 - English	0.20	0.28	0.84	0.93
P23 - French	0.19	0.26	0.78	0.89
P23 - Japanese	0.18	0.25	0.84	0.91
Mean	0.21	0.29	0.83	0.90

Tables show detailed results for POLQA estimated using a 300 ms temporal window

Mean Absolute Difference (MAD)	$\text{MAD} = \sum_{n=1}^{N} \frac{1}{n}(\lvert E(n)\rvert)$
Root Mean Square Error (RMSE)	$\text{RMSE} = \sqrt{\sum_{n=1}^{N} \frac{1}{n}(\lvert E(n)\rvert^2)}$
Pearson Correlation Coefficient (R)	Measures the dependence between P_e and P_t and takes a value in the range $[-1, 1]$.
F1 Score (F1)	$\text{F1} = \frac{2TP}{2TP + FP + FN}$, where TP is the true positive rate, FP is the false positive rate and FN is the false negative rate.

Conclusion

- This paper proposed: 1) the first short-time, non-intrusive POLQA estimator, NISA, and 2) a novel compressed representation of modulation features and MFCCs
- The results demonstrate that in terms of relative RMSE, with a 300 ms context, the proposed method outperforms the baseline method by 38.3%
- Joint short time prediction of VAD estimates and POLQA with an average estimation accuracy of 0.29 POLQA (RMSE) and VAD F1 score of 0.90 across test sets that included unseen languages