2021 Thirteenth International Conference on Quality of Multimedia Experience (QoMEX)

Full-Reference and No-Reference Objective Evaluation of Deep Neural Network Speech

Stephen Voran
Institute for Telecommunication Sciences
Boulder, Colorado, USA
svoran@ntia.gov

***Abstract*—Objective speech quality and intelligibility estimators do not correctly assess speech generated by deep neural networks (DNNs). We use 256 speech files and subjective scores that cover 14 DNN speech conditions and 18 nonDNN speech conditions to show that 8 different full-reference (FR) estimators consistently underestimate subjective scores for the DNN conditions. Conversely, we find that five no-reference (NR) estimators consistently overestimate subjective scores for the DNN conditions. We show that a rudimentary but effective solution to these shortcomings is to simply average an FR result with an NR result. We also explore root causes and propose more fundamental solutions. It has been previously suggested that FR estimators over-penalize inaudible timing variations or jitter. We conduct several experiments that measure and remove jitter from spectral representations of DNN speech inside FR estimators. Jitter removal compensates for some of the underestimation, thus confirming that jitter is a part of the cause. In additional experiments we show that power mismatches on a syllabic time-scale also contribute to the underestimation issue in FR estimators. Regarding NR estimators, we suggest that they can be trained to accurately rate DNN speech when sufficient speech signals and corresponding subjective scores are available.**

***Index Terms*—DNN speech, neural speech, objective estimator, speech intelligibility, speech quality**

I. Introduction

Speech signals are essential in many multimedia experiences and the quality or intelligibility of speech signals can very strongly influence the overall multimedia quality of experience. Objective estimators are can provide efficient and effective means for determining speech quality and intelligibility in many existing multimedia telecommunications environments. Full-reference (FR) estimators compare input and output speech signals of a system-under-test. No-reference (NR) estimators require only the output. Now Deep Neural Networks (DNNs) that generate speech from text or from speech encoder parameters are emerging, thus enabling DNN-based speech coding for multimedia telecommunications applications [1]–[12].

But measuring DNN speech is problematic for both FR and NR speech quality and intelligibility estimators. Consequently, evaluations of DNN speech are best done by relatively slow and expensive subjective tests. In [6] it was noted that the well-established FR quality estimator POLQA [13] "did not reflect informal listening impressions" for speech produced by the parametric WaveNet DNN coder. The objective estimates were far lower than the listening impressions. This finding has also been stated in [8], [9]. In [14] a much more detailed study identifies aspects of speech signals that may contribute to the issue. These include "micro-alignment time shifts" and "pitch accuracy." These are related to each other because warping or distorting the time axis will also distort pitches. They are also consistent with the statement in [6] that artificially low quality estimates were "not unexpected as the parametric WaveNet coder changes the signal waveform and the timing of the phones."

In the next section we describe the speech files and subjective scores we use to study this issue. We then report and discuss how objective estimators respond to this data, and resolutions to shortcomings in those responses. In Section IV, we detail our experiments with modifying FR estimators to achieve the invariances needed to more accurately evaluate DNN speech.

II. Speech Files and Subjective Scores

We received crowd-sourced subjective test scores and the corresponding wideband speech files ($f_s = 16k$) associated with the five tests described in [9]–[11]. These tests followed the multi-stimulus test with hidden reference and anchor (MUSHRA) [15] paradigm where listeners are instructed to submit a "subjective judgement of the quality level for each of the sound excerpts" using a 0 to 100 scale. This scale is divided into five equal intervals that are additionally labeled "bad," "poor," fair," "good," and "excellent" [15]. Each speech file received ratings from 100 different listeners. The resulting scores span most of the 100 point scale. The lowest per-file mean score is 10.8 and the highest is 100.0. The largest 95% confidence interval (CI) for a file is 4.4 and the average CI is 2.7, so the 100 point scale is well-resolved.

Four of the tests covered six conditions each, and the fifth test covered eight conditions. (A "condition" is either a speech codec or original speech.) Thus 32 different conditions were tested. Eight speech files were used to test each condition, and these files were different in the different tests. With 32 conditions and 8 files per condition, we have 256 speech files and subjective scores to drive our study.

The five tests evaluate 14 DNN-based codecs and 18 conventional (nonDNN) conditions in total. The 14 DNN codecs include variations of LPCNet [9]–[11], WaveRNN+ [9], and

WaveNet [11]. The 18 nonDNN conditions serve as anchors or points of reference in the tests. These include original (appears 5 times), μ-law PCM, Opus 9 kbps (appears 4 times), Opus 6 kbps (appears twice), Speex 4 kbps (appears 4 times), and MELP 2.4 kbps (appears twice).

We listened to the speech files and found that the speech produced by WaveRNN+ per [9] and LPCNet per [9]–[11] includes some raspiness, or speech-correlated noise that sounds similar to quantization noise but with a bit more spectral or temporal structure than quantization noise. This impairment decreases as the number of dense equivalent units is increased [9] or when quantization is removed [10]. Speech from WaveNet per [11] does not have this artifact, but sounds as if some syllables are over- or under-emphasized in a slightly frequency-selective manner. The result is a slightly unnatural sound, as if the talker suffers from a mild impairment or is making a poor attempt to add dramatic flair to the delivery.

III. Objective Evaluations

MUSHRA is a quality scale (it quantifies the pleasing or non-pleasing nature of speech sounds) which is different from, yet related to, an intelligibly scale (which quantifies information transferred by speech signals). To maximize the breadth of our work, we applied both quality and intelligibility estimators to the speech files described in Section II. The nine quality and four intelligibility estimators are labeled numerically in Table I. The first eight estimators are FR and the next five are NR.

FR estimators use "reference speech" (the input to the speech encoder or the system-under-test) as well as "test speech" (the output of the speech decoder or the system-under-test). NR estimation is more challenging because reference speech is not used, and this also means that NR estimation is possible in situations where FR is not. The final entry in Table I is the average of two others and the motivation for this is provided later in this section.

Note that fixed time offsets (delays) between reference and test files do not influence subjective scores and such offsets must be removed as part of the FR estimation process. Some FR implementations do this internally and others require it be done externally. Thus we used delay estimates from WB-PESQ to time-align reference and test signals before processing by the other FR estimators. Also note that FR estimators 1, 2, 3, and 4 attempt to track varying delays between reference and test files and the others do not.

We used a quadratic function to map each estimator output to the MUSHRA scale (0 to 100),

$$c_0 + c_1 q + c_2 q^2 = \hat{M} \approx M, \tag{1}$$

where q is an estimator output, M is a MUSHRA score, and the coefficients c_i minimize the RMS value of $\hat{M} - M$ over the 18 nonDNN conditions under the constraint that (1) is monotonic over the entire range of the data. These 18 conditions have been broadly available and would have been used in the development of at least some of the estimators. Fig. 1 shows example results for five estimators.

TABLE I
Objective estimators and labels used in Figs. 2 and 3.

Label	Name and Reference	Type	Dimension
1	POLQA v2.4 [13]	FR	Quality
2	WB–PESQ [16]	FR	Quality
3	ViSQOL V3 [17]	FR	Quality
4	ViSQOL [18]	FR	Quality
5	PEMO [19]	FR	Quality
6	SIIB^{Gauss} [20]	FR	Intelligibility
7	ESTOI [21]	FR	Intelligibility
8	STOI [22]	FR	Intelligibility
9	NISQA [23]	NR	Quality
10	WAWEnet mode 1 [24]	NR	Quality
11	WAWEnet mode 2 [24]	NR	Quality
12	WAWEnet mode 3 [24]	NR	Quality
13	WAWEnet mode 4 [24]	NR	Intelligibility
14	(WB–PESQ + WAWEnet mode 1) / 2	FR+ NR	Quality

A. Objective-Subjective Errors and Agreement

The plots in Fig. 1 show a fairly high level of agreement for the nonDNN conditions and consistent error trends for many of the DNN conditions. Fig. 2 shows that these trends are borne out across all 13 estimators. The FR estimators (1-8) underrate the DNN conditions by 20 to 31 MUSHRA points and the NR estimators (9-13) overrate them by 17 to 27 points. Across the 14 estimators, the RMSE for DNN conditions ranges from 20 to 34, and these values are consistently larger than the RMSE values for nonDNN conditions, which range from 5 to 14.

These results show that DNN speech is more challenging for all 13 estimators. One might attempt to address this by refitting (1) using both nonDNN and DNN conditions. Fig. 3 shows the results. As expected, the total error is more equitably distributed between the DNN and nonDNN conditions but simple fitting cannot address the root issue. Fig. 3 shows the magnitude of the mean error is greater for DNN than nonDNN conditions in all 13 cases. The RMSE for the DNN conditions exceeds that of the nonDNN conditions for FR estimators, but they are more nearly matched for NR estimators.

The complementary nature of the FR and NR errors suggests a rudimentary but pragmatic solution — allow FR and NR estimators to work together to form a new FR estimator. We found averaging the output of WB-PESQ with the output of WAWEnet mode 1 produced the smallest errors overall so that combination is "Estimator 14" in Table I and in Figs. 2 and 3. This hybrid solution gives very small mean errors for both DNN and nonDNN conditions and an overall RMS error of 10 points. In spite of this easy and convenient solution, we continue to investigate how to make individual FR and NR estimators that more accurately evaluate DNN speech.

Figs. 2 and 3 show that in terms of per-condition RMSE, the NR estimators are as good as the FR ones — the ranges covered by the FR and NR estimators differ by no more than 1 point in either figure. These figures also show that the RMSEs of the intelligibility estimators fall inside the RMSE range defined by the quality estimators. Overall, quality estimators and intelligibility estimators do equally well at estimating MUSHRA quality scores for these conditions.

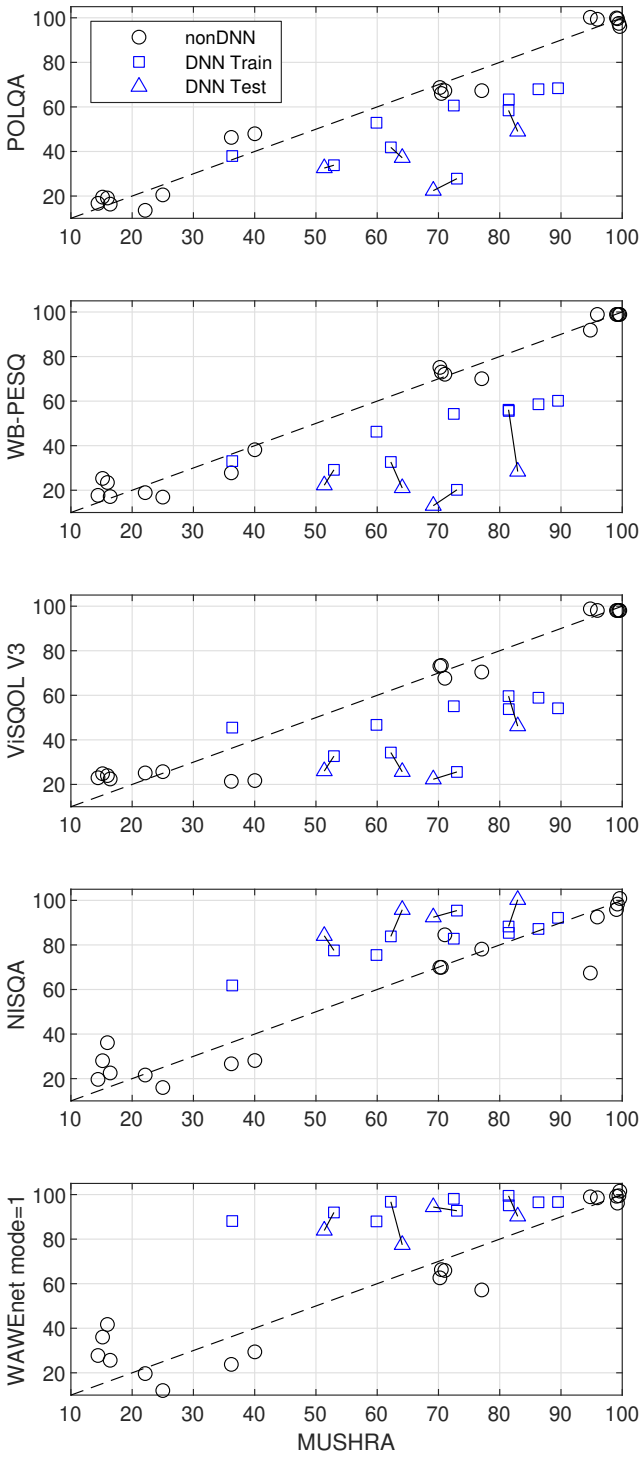

Fig. 1. Example scatter plots of fitted objective estimates $\hat{M}$ produced by (1) vs MUSHRA scores for 3 FR and 2 NR estimators applied to 14 DNN conditions (blue) and 18 nonDNN conditions (black). When a DNN condition has both a training case and a testing case, data for the two cases are connected by a black line to allow convenient comparisons.

Four of the DNN codecs were evaluated a second time to assess robustness to speech from outside the training database. In Fig. 1 these train and test conditions are indicated by squares and triangles respectively, and are connected by lines. The MUSHRA scores for these condition pairs are generally very close and this shows the codecs are robust to speech outside the training speech. But the objective estimates often show much larger spreads, falsely implying a lack of codec robustness to speech outside the training speech.

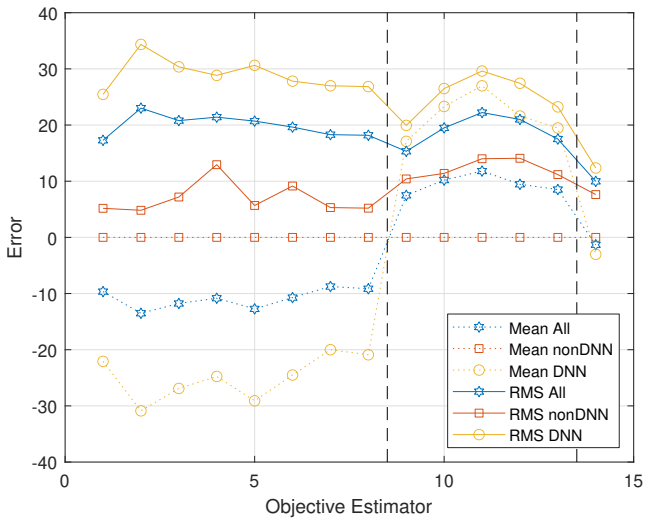

Fig. 2. Per-condition mean error (dotted) and RMS error (solid) for 14 estimators identified in Table I. Estimators have been fitted to the MUSHRA scale using 18 nonDNN conditions. Dashed vertical lines emphasize the contrasting results for FR estimators (1-8), NR estimators (9-13) and the experimental hybrid estimator (14).

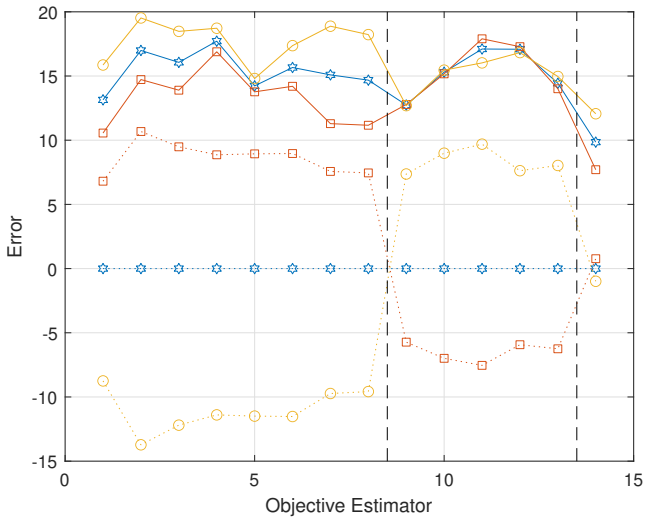

Fig. 3. Per-condition mean error (dotted) and RMS error (solid) for 14 estimators identified in Table I. Estimators have been fitted to the MUSHRA scale using all 32 conditions. Dashed vertical lines emphasize the contrasting results for FR estimators (1-8), NR estimators (9-13) and the experimental hybrid estimator (14). Legend of Fig. 2 applies in this figure as well.

B. Expanding Estimator Scope to include DNN Speech

The NR estimators are DNN-based and they have learned the relationships between mel-frequency log energy features and speech quality [23] or unprocessed speech waveform samples and quality and intelligibility [24]. The DNN speech signals used here exhibit the properties that these NR estima-

tors associate with high quality and intelligibility — so much so that DNN speech is generally overrated. We expect that NR estimators could be trained to accurately rate DNN speech if sufficient speech signals and subjective scores were available.

The FR estimators compare perceptually-motivated spectral representations of reference and test signals in a manner that seeks to emulate human cognition. The DNN conditions produce spectral representations that often register as very different inside an estimator, when in fact the test speech sounds similar to the reference speech. This may be resolved by augmenting the estimator to have an appropriate amount of invariance to the differences that are unique to DNN speech. If this augmentation does not interfere with the existing modeling, then the estimator may work well for both DNN and nonDNN speech.

The DNN conditions were not seen by the 13 estimators when the estimators were developed and the estimation errors reported here are not indications of failure. Rather they simply confirm that DNN speech contains fundamentally new impairments that are out-of-scope for the 13 estimators.

The issue of scope is a reoccurring one for objective estimators. An ideal estimator would perfectly model average human audition and cognition and would thus produce results that agree with average human ratings for all existing and future telecommunication technologies. But a practical estimator will only contain sufficient modeling to produce accurate results for a reasonably broad collection of existing telecommunications technologies.

As telecommunications evolves, entirely new classes of impairments emerge and a practical estimator may then fail to agree with human ratings in the presence of these new impairments. When this happens we are motivated to augment existing modeling to more closely emulate the human response to the new impairments and thus expand the estimator scope to include the new impairments.

This cycle played out multiple times as telecommunications evolved from analog to digital waveform coding to digital model-based coding, from circuit-switched to packet-switched transport, from wired to wireless transmission, and from narrowband speech to wideband speech and beyond [25]–[28]. It is not surprising that the advent of DNN-based speech coding necessitates another of these development and evaluation cycles.

IV. Adding Invariances to Estimators

We performed several experiments to determine if additional invariances could reduce FR estimation error on DNN speech conditions. We selected four DNN conditions and one nonDNN condition from the same region of the MUSHRA scale (69 to 83) to allow a focus on prominent examples of the issue. Our initial experiments used ViSQOL [18] since MATLAB®[1] ViSQOL code allows for easy experimentation. The selected conditions and the ViSQOL estimation errors $\hat{M} - M$ ($\hat{M}$ fit on the 18 nonDNN conditions) are shown in columns 1 and 6 of Table II.

[1]Disclaimer: The commercial software MATLAB® was used to perform this analysis for the convenience of the author; such use does not imply recommendation or endorsement by the National Telecommunications and Information Administration, nor does it imply that the software used is necessarily the best available for the particular application or uses.

A. Removing Jitter

We plotted input and output waveforms for the DNN conditions and compared them. We found that pitch pulse locations did not maintain consistent relationships between input and output waveforms but instead exhibited continually evolving relationships. This provided initial confirmation of timing variations (jitter) in the four selected DNN conditions, as previously suggested in [6] and [14]. In the WaveNet conditions, comparison of input and output waveforms confirmed power mismatches on a syllabic time-scale as reported in Section II. We theorized that jitter and power matching were more significant to ViSQOL than to listeners, thus contributing to the underrating of the DNN conditions. To test this theory we modified ViSQOL to increase its invariance to these two impairments.

ViSQOL uses NSIM [29] to compare salient time-frequency patches extracted from critical band dB-scale spectral representations of the reference and test signals. For wideband speech these 21-band representations are computed from 32 ms long frames spaced every 16 ms (256 samples) and are represented here by X_{ij} (reference) and Y_{ij} (test), where $i \in \{1, 2, ..., 21\}$ and $j \in \{1, 2, ..., N\}$ are the frequency band and frame indices, respectively. We decreased the 16 ms spacing by a factor of $R = 128$ to 125 μs (2 samples) to produce a high time-resolution 21-band spectral representation Z_{ik} with $k \in \{1, 2, ..., RN\}$ for the test signal only. Next we selectively sub-sampled the sequence of Z_{ik} to create a standard resolution dejittered sequence Y_{ij} for ViSQOL to compare with the sequence X_{ij} in standard fashion:

$$\begin{aligned} Y_{\cdot j} &= Z_{\cdot g(j)} \\ &= f(X_{\cdot j}, [Z_{\cdot L(j,R)}, Z_{\cdot L(j,R)+1}, \ldots, Z_{\cdot U(j,R)}]). \end{aligned} \quad (2)$$

The "$\cdot$" symbol indicates that (2) operates on all bands at once. The function f compares the j^{th} frame of X ($X_{\cdot j}$) with a range of candidate frames in Z and extracts the frame with the largest correlation (across the 21 bands) to $X_{\cdot j}$. The frame index of the extracted frame is $g(j)$. The range of candidate frames is set by the lower limit function L and the upper limit function U which initially define a +/- 64 ms search range. That range is narrowed to enforce temporal monotonicity, $g(j) < g(j+1)$, and to prevent searching outside the range $[1, RN]$. This means L is also a function of $g(j-1)$ and U is also a function of N (not shown for clarity). Since the number of candidate frames is variable, they are passed into f as the columns of an array.

This process produces a dejittered Y and the frame index history $g(j)$. If there is no jitter then $g(j) = R(j-1)+1, \forall j$. So $g(j) - (R(j-1)+1)$ is a measure of jitter and we report the RMS value of this (averaged over the 8 files for each condition) in column 2 of Table II. Pitch errors are related to jitter. We used the cepstrum method of the MATLAB pitch

TABLE II

SPEECH SIGNAL PROPERTIES MEASURED FOR FIVE CONDITIONS (MEAN OVER EIGHT SPEECH FILES PER CONDITION), ESTIMATION ERRORS, AND CHANGES IN ESTIMATES INDUCED BY DEJITTERING ALONE AND DEJITTERING WITH POWER MATCHING. WB-PESQ INDICATED BY "PESQ" DUE TO SPACE LIMITATIONS.

Condition	**RMS jitter (ms)**	**RMS pitch error (Hz)**	**Relative pitch error magnitude (%)**	**Power mismatch (dB)**	**Estimation error $\hat{M} - M$ (MUSHRA)**		**Change in $\hat{M}$ due to dejitter (MUSHRA)**		**Change in $\hat{M}$ due to dejitter + power match (MUSHRA)**	
					ViSQOL	***PESQ***	***ViSQOL***	***PESQ***	***ViSQOL***	***PESQ***
Opus (9 kbps) [11]	12.3	10.6	2.3	3.6	-11	+2	0	0	+1	+4
WaveNet Train [11]	18.9	13.3	4.6	7.4	-42	-53	+22	+10	+25	+18
WaveNet Test [11]	19.7	17.9	6.0	7.1	-54	-56	+31	+8	+33	+14
LPCNet Train [10]	13.8	12.5	3.2	3.2	-28	-25	+7	+8	+8	+14
LPCNet Test [10]	16.5	17.0	6.8	3.9	-36	-55	+11	+12	+12	+20

function to measure pitch in reference and test speech signals. The RMS pitch error is given in column 3 of Table II. The mean magnitude of the pitch errors, expressed as a percent of the reference pitch appears in column 4.

B. Matching Power

We also experimented with matching the total power of each pair of frames (reference and test) after dejittering to eliminate the observed power mismatches:

$$\tilde{Y}_{\cdot j} = Y_{\cdot j} + p(j), \quad p(j) = 10log_{10}\left(\frac{\sigma_x^2(j)}{\sigma_y^2(j)}\right),$$
$$\sigma_x^2(j) = \sum_{i=1}^{21} 10^{\frac{X_{ij}}{10}}, \quad \sigma_y^2(j) = \sum_{i=1}^{21} 10^{\frac{Y_{ij}}{10}}. \qquad (3)$$

The sequence of power matching values $p(j)$ (in dB) is a measure of the power adjustments made and is also a measure of the power mismatch before the adjustments. The RMS value of this sequence of dB values is given in column 5 of Table II. Columns 2 and 5 in Table II show speech signal measurements in the ViSQOL frame domain. In the WB-PESQ frame domain the values are different but the trends are clearly preserved.

C. Invariances Reduce Estimation Errors

Column 6 of Table II reports the average ViSQOL and WB-PESQ estimation errors for the five conditions. As expected, these FR estimation errors are large and negative for the DNN conditions and much smaller for the nonDNN (Opus 9 kbps) condition.

We used analogous steps to remove jitter and match power in ViSQOL, WB-PESQ, and ESTOI. The rightmost columns of Table II show the changes in the final MUSHRA estimates of ViSQOL and WB-PESQ due to dejittering alone (column 7) and due to dejittering followed by power matching (column 8). In the case of ESTOI we observed similar estimation errors but much smaller score increases from dejittering and power matching. Those increases never exceeded 5 points.

As expected, Table II shows that larger jitter is generally associated with larger pitch errors and that these are also associated with larger score improvements when jitter is removed, consistent with the theory that inaudible jitter causes score underestimation. That table also shows that increased power matching is often associated with larger score increases, indicating that power mismatch may also contribute to underestimation. Jitter removal contributes the majority of the score increase, and power matching contributes a lesser portion. The testing conditions show more jitter, and their scores improve more than the training conditions. For all three estimators, the two new invariances increase the DNN scores dramatically more than they increase the nonDNN (Opus 9 kbps) score. This indicates that these two new invariances can indeed contribute to the solution of the problem.

We have used available data to gain initial insights. One could also create specific additional data to enable a more thorough investigation. That is, in a future study one might add simulated jitter to speech at controlled levels and compare the responses of the original and modified objective estimators with subjective scores.

D. Discussion

The contributions are variable between ViSQOL, WB-PESQ and ESTOI, and they do not completely solve the issue for any of these three. We attribute this to the fact that these are the results of relatively simple experiments where we patched minimally-invasive modifications onto the early processing stages of FR estimators. These are not fully integrated, optimized, and calibrated enhancements. Additionally, in spite of the high time-resolution spectral representations, the native frame sizes (32 ms in ViSQOL and WB-PESQ or 25.6 ms in ESTOI) inhibit the full removal of jitter at timescales shorter than these values. The comparison stages of the various FR estimators are complex and they make comparisons in light of local time-frequency contexts established by the signals themselves. To be fully effective, any new invariances must be properly integrated with them, not blindly prepended to them. And all modifications must be fully tested across a much wider range of codecs, channel conditions, noise environments, and other conditions to identify and resolve any negative side-effects.

In general it will not be desirable to remove unlimited amounts of jitter or power fluctuation without reflecting at least some of this correction by lowering the output score. The audibility of these impairments will depend on their magnitudes and temporal distributions and these should be considered in relationship to the speech.

In a separate but related experiment we applied smoothed power matching to the WaveNet files in the time-domain and heard a clear improvement in quality and naturalness. This required access to the reference speech and thus would not be a practical enhancement strategy, but it does indicate an area where there is room for improvement and it also reaffirms that power matching adjustments should be reflected in objective estimates.

V. Conclusions

We have provided broad novel evidence confirming that FR objective speech quality and intelligibility estimators significantly underestimate subjective scores of DNN speech. We have also documented a consistent trend for current NR estimators to overestimate subjective scores of DNN speech.

A simple combination of FR and NR estimates is a solution, but it is not as satisfying as one that addresses the root causes. Our experiments clearly demonstrate that inaudible jitter is indeed a strong factor in the FR speech quality estimators ViSQOL and WB-PESQ, and that power mismatches contribute in a lesser way. Additional training should resolve the problem in DNN-based NR estimators when sufficient speech files and scores are available.

Thus we have identified multiple paths toward new FR and NR estimators that more accurately evaluate the DNN speech signals that will be appearing in multimedia telecommunications. Such estimators might also be leveraged in automated training and evaluation cycles leading to more rapid improvements in DNN speech quality, intelligibility, and naturalness.

Acknowledgment

Sincere thanks to Dr. Jan Skoglund of Google. This work is possible only due to speech files and subjective scores that Jan provided.

References

[1] M. Cernak, A. Lazaridis, A. Asaei, and P. N. Garner, "Composition of deep and spiking neural networks for very low bit rate speech coding," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 24, no. 12, pp. 2301–2312, 2016.

[2] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "WaveNet: A generative model for raw audio," in *Proc. 9th ISCA Speech Synthesis Workshop*, 2016.

[3] S. Mehri, K. Kumar, I. Gulrajani, R. Kumar, S. Jain, J. Sotelo, A. Courville, and Y. Bengio, "SampleRNN: An unconditional end-to-end neural audio generation model," arXiv:1612.07837, Dec. 2016.

[4] A. Tamamori, T. Hayashi, K. Kobayashi, K. Takeda, and T. Toda, "Speaker-dependent WaveNet vocoder," in *Proc. Interspeech*, 2017.

[5] N. Kalchbrenner, E. Elsen, K. Simonyan, S. Noury, N. Casagrande, E. Lockhart, F. Stimberg, A. van den Oord, S. Dieleman, and K. Kavukcuoglu, "Efficient neural audio synthesis," in *Proc. 35th Intl. Conf. on Machine Learning*, 2018, pp. 2410–2419.

[6] W. Kleijn, F. Lim, A. Luebs, J. Skoglund, F. Stimberg, Q. Wang, and T. Walters, "Wavenet based low rate speech coding," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing*, 2018, pp. 676–680.

[7] C. Gârbacea, A. van den Oord, Y. Li, F. C. Lim, A. Luebs, O. Vinyals, and T. Walters, "Low bit-rate speech coding with VQ-VAE and a WaveNet decoder," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing*, 2019, pp. 735–739.

[8] J. Klejsa, P. Hedelin, C. Zhou, R. Fejgin, and L. Villemoes, "High-quality speech coding with SampleRNN," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing*, 2019, pp. 7155–7159.

[9] J. Valin and J. Skoglund, "LPCNet: Improving neural speech synthesis through linear prediction," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing*, 2019, pp. 5891–5895.

[10] J. Valin and J. Skoglund, "A real-time wideband neural vocoder at 1.6kb/s using LPCNet," in *Proc. Interspeech*, 2019, pp. 3406–3410.

[11] J. Skoglund and J. Valin, "Improving Opus low bit rate quality with neural speech synthesis," in *Proc. Interspeech*, 2020.

[12] F. Lim, W. Kleijn, M. Chinen, and J. Skoglund, "Robust low rate speech coding based on cloned networks and WaveNet," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing*, 2020, pp. 6769–6773.

[13] *Recommendation ITU-T P.863, "Perceptual objective listening quality prediction," International Telecommunication Union, Geneva, 2018.*

[14] W. Jassim, J. Skoglund, M. Chinen, and A. Hines, "Speech quality factors for traditional and neural-based low bit rate vocoders," in *Proc. Twelfth Intl. Conf. on Quality of Multimedia Experience*, 2020.

[15] *Recommendation ITU-R BS.1534-3, "Method for the subjective assessment of intermediate quality level of audio systems," International Telecommunication Union, Geneva, 2015.*

[16] *Recommendation ITU-T P.862.2, "Wideband extension to Recommendation P.862 for the assessment of wideband telephone networks and speech codecs," International Telecommunication Union, Geneva, 2007.*

[17] M. Chinen, F. Lim, J. Skoglund, N. Gureev, F. O'Gorman, and A. Hines, "ViSQOL v3: An open source production ready objective speech and audio metric," in *Proc. Twelfth Intl. Conf. on Quality of Multimedia Experience*, 2020. Code at: https://github.com/google/visqol

[18] A. Kokaram, A. Hines, J. Skoglund, and N. Harte, "ViSQOL: An objective speech quality model," *EURASIP Journal on Audio, Speech, and Music Processing*, 2015. Code at http://www.mee.tcd.ie/~sigmedia/Resources/ViSQOL

[19] V. Emiya, E. Vincent, N. Harlander, and V. Hohmann, "Subjective and objective quality assessment of audio source separation," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2046–2057, 2011. Code at: http://bass-db.gforge.inria.fr/peass/

[20] S. Van Kuyk, W. Kleijn, and R. Hendriks, "An instrumental intelligibility metric based on information theory," *IEEE Signal Processing Letters*, vol. 25, no. 1, pp. 115–119, 2018. Code at: https://stevenvankuyk.com/matlab_code/

[21] J. Jensen and C. H. Taal, "An algorithm for predicting the intelligibility of speech masked by modulated noise maskers," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 24, no. 11, pp. 2009–2022, 2016. Code at: http://kom.aau.dk/ ∼ jje/

[22] C. Taal, R. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2125–2136, 2011. Code at: http://ceestaal.nl/code/

[23] G. Mittag and S. Möller, "Non-intrusive speech quality assessment for super-wideband speech communication networks," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing*, 2019, pp. 7125–7129. Code at: https://github.com/gabrielmittag/NISQA

[24] A. Catellier and S. Voran, "WAWEnets: A no-reference convolutional waveform-based approach to estimating narrowband and wideband speech quality," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing*, 2020, pp. 331–335. Code at: https://github.com/NTIA/WEnets

[25] J. Beerends and J. Stemerdink, "A perceptual speech-quality measure based on a psychoacoustic sound representation," *J. Audio Engineering Society*, vol. 42, pp. 115–123, March 1994.

[26] S. Voran, "Objective estimation of perceived speech quality. I. Development of the measuring normalizing block technique," *IEEE Trans. on Speech and Audio Processing*, vol. 7, no. 4, pp. 371–382, July 1999.

[27] *Recommendation ITU-T P.862, "Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs," International Telecommunication Union, Geneva, 2001.*

[28] S. Voran, "A bottom-up algorithm for estimating time-varying delays in coded speech," in *Proc. Third Intl. Conf. on Measurement of Speech and Audio Quality in Networks*, May 2004.

[29] A. Hines and N. Harte, "Speech intelligibility prediction using a neurogram similarity index measure," *Speech Communication*, vol. 54, 2012.