Audio Similarity is Unreliable as a Proxy for Audio Quality

Pranay Manocha1, Zeyu Jin2, Adam Finkelstein1
1Department of Computer Science, Princeton University, USA
2Adobe Research, USA
{pmanocha,af}@cs.princeton.edu, zejin@adobe.com

Abstract

Many audio processing tasks require perceptual assessment. However, the time and expense of obtaining "gold standard" human judgments limit the availability of such data. Most applications incorporate full reference or other similarity-based metrics (e.g. PESQ) that depend on a clean reference. Researchers have relied on such metrics to evaluate and compare various proposed methods, often concluding that small, measured differences imply one is more effective than another. This paper demonstrates several practical scenarios where similarity metrics fail to agree with human perception, because they: (1) vary with clean references; (2) rely on attributes that humans factor out when considering quality; and (3) are sensitive to imperceptible signal level differences. In those scenarios, we show that no-reference metrics do not suffer from such shortcomings and correlate better with human perception. We conclude therefore that similarity serves as an unreliable proxy for audio quality.

Index Terms: audio quality, speech quality, similarity metrics, perceptual metric, speech enhancement

1. Introduction

Speech quality assessment (SQA) plays a critical role in a range of audio and speech applications, including telephony, VoIP, hearing aids, automatic speech recognition, and speech enhancement. The "gold standard" for SQA involves human listening tests. However, these subjective evaluations are time consuming and expensive, as they require a human in the loop and often need to be repeated many times for every recording. Thus automatic (objective) SQA methods are often more practical, and roughly fall into two categories – whether or not they rely on a reference recording.

Full-reference metrics are also known as intrusive or similarity metrics (e.g. PESQ [1], POLQA [2], VISQOL [3], DPAM [4], CDPM [5] and others [6–8]). They require a clean reference to which a corrupted signal can be compared as the basis for a quality rating. On the other hand, no-reference metrics, also called non-intrusive [9] metrics, (e.g. DNSMOS [10], NISQA [11], SQAPP [12] and others [13–17]) are designed to output a rating on an absolute scale, without access to a clean reference. Researchers commonly rely on full-reference metrics as a proxy for audio quality, because they were introduced earlier – consider, e.g. SNR. One of the most impactful is PESQ [1], introduced decades ago for telephony and still used today for enhancement [4, 5, 18–22], vocoders [23], and transmission codecs [24, 25]. PESQ correlates well with subjective listening tests [10, 26], and PESQ labels have also been leveraged for training differentiable quality metrics [7, 13, 27]. Researchers rely on objective metrics like PESQ for comparing the effectiveness of various methods, sometimes reporting small gains (~0.1 PESQ). However, PESQ has acknowledged shortcomings [29, 30], and may not be reliable to detect subtle differences [4]. Researchers have explored ways to ameliorate such shortcomings and improve robustness of such objective metrics [2, 4]. However, this paper describes experiments suggesting that the inherent problem may be the overall formulation relying on a clean reference. For example, Figure 1 sketches a scenario where two different "clean" reference recordings are used to evaluate the quality of a test recording. An effective metric should report the same quality for the test, as long as the reference is “clean” – but similarity metrics naturally report different quality measures in this scenario, because the recording setup of the test happens to be acoustically closer to one of them.

This paper investigates such limitations of similarity metrics, with the goal of informing future SQA research. We describe several scenarios wherein we empirically evaluate seven similarity metrics (L1, L2, Multi-resolution STFT, PESQ, VISQOL, DPAM and CDPM), and show that these contrast with subjective quality judgments. Overall, we find that: (1) similarity metrics fail to capture the multi-modality of audio quality relative to "clean" recordings made in different environments; (2) mismatched training and evaluation datasets cause objective ratings to differ from subjective ratings; and (3) similarity metrics emphasize imperceptible differences, and therefore fail for recordings that sound perceptually indistinguishable but is sampled from a different distribution. Moreover, we also show specific examples (e.g. high-pass filtering or comparing cross-perturbations) where similarity metrics contrast with subjective judgments. We also evaluate four no-reference metrics (DNSMOS, NISQA, SQAPP and NORESQQA) on those scenarios, and show that they do not suffer from the same shortcomings. These findings illustrate some weaknesses of similarity metrics (e.g. PESQ), and help inform researchers on how best to use them across various tasks. Listening examples are available here:
https://pml.cs.princeton.edu/pub/Manocha_2022_ASI/
Figure 2: Scenario 2: Test 1 and Test 2 are two recordings from models trained on two different datasets (DAPS [31] and VCTK [32]) respectively. Test 1 has a higher quality than the Test 2 recording as confirmed by listening tests. When compared across a reference that lies acoustically closer to Test 2 \((s_2 < s_1)\), we see that it may be more similar to Test 2, even though Test 1 has higher quality.

Figure 3: Scenario 3: Test 1a and Test 2 are two recordings from models trained on different datasets (DAPS [31] and VCTK [32] respectively). Test 1b is created by acoustically matching recordings from Test 1a to Test 2. Even though recordings from Tests 1b and 2 sounds perceptually indistinguishable, they may have different similarities \((s_1 \text{ and } s_2)\) with a reference, which suggests these metrics emphasize imperceptible differences.

2.2. One reference with two test recordings

Figure 2 illustrates a scenario in which we hypothesize that similarity and quality can have negative correlation – comparing two signals output from different processes with the single reference. Our experiment involves speech enhancement (SE), whose goal is to improve the perceptual quality of speech signals by removing background noise and other perturbations. Because subjective evaluations are not comparable across different experimental settings [33], researchers rely on objective metrics (e.g. PESQ for quality and STOI [34] for intelligibility) on a common dataset (e.g. VCTK [32]) to compare their models.

Here, we also use the widely accepted VCTK evaluation dataset \((X)\) as a clean reference. We use the HiFi-GAN [18] architecture to train enhancement models on two datasets: \(M_1\) trained on DAPS [31] and \(M_2\) trained on VCTK [32]. Recordings from the VCTK set \((X)\) are passed through the two models to produce \(Y_1 = M_1(X)\) and \(Y_2 = M_2(X)\). We then compare \(Y_1\) and \(Y_2\) across various objective measures to assess performance. Finally, we do an MOS listening study to show how well different objective metrics correlate with subjective ratings. The results are summarized in Section 3.2.

2.3. Matching datasets acoustically

Figure 3 illustrates a scenario wherein similarity metrics over-emphasize imperceptible differences such as frequency energy, inaudible noise and phase differences. To obtain parallel recordings coming from two different distributions, we start with the recordings \(Y_1\) and \(Y_2\) from Section 2.2. Next we process the output samples generated by DAPS model \(Y_1\) so as to match the acoustics of the VCTK model \(Y_2\), until they are almost perceptually indistinguishable, as follows.

We first apply low- and high-pass filtering to remove unwanted frequency components, and prevent aliasing. Next we perform various steps, in order, based on the observation that \(Y_1\) and \(Y_2\) have different acoustics and background noise: (1) Equalization (EQ) matching following Germain et al. [35] equalizes timbral content and background noise between audio recording environments. (2) Breath removal reduces perceptual differences by eliminating the minor breath sounds present in VCTK.

(3a) Per-frequency energy normalization [36] equalizes the energies at corresponding frequencies across samples. Because changing frequency energy may produce a “frequency leak” artifact if phase remains unchanged, we also experimented using 0, 200 and 1000 iterations of Griffin-Lim to refine the phase. (3b) Per-frequency energy normalization with ground-truth clean phase. The results are summarized in Section 3.3.
### 3. Results

#### 3.1. Two reference with one test recording

Refer to Tables 1 and 2 for details. First, we see whether recordings from *studio2* and *studio3* are rated as equally good, or not. Table 1 shows the scores of a subjective listening study conducted on Amazon Mechanical Turk (AMT), where each subject is asked to rate the sound quality of an audio snippet on a scale of 1 to 5, with 1=Bad, 5=Excellent. Overall, we collect 1046 ratings per condition from 244 unique native English speakers. The workers have to pass a hearing test where they report all objective measures and subjective ratings - Table 1 - row1) for training an enhancement model leads to a better model with higher audio quality (Table 3 last row). However, if using similarity metric as proxy for quality, higher-quality models may be labelled as lower quality.

#### 3.3. Matching datasets acoustically

Refer to Table 4. Recall that we developed a set of pre-processing stages to acoustically match the recordings from the DAPS trained model to the VCTK trained model used in Section 3.2. For all stages, we report all objective metrics, and subjective listener ratings. The subjective listening tests are conducted on AMT, where a subject is asked to rate the quality of an audio snippet on a scale of 1 to 5, collecting around 1974 ratings per condition from 371 unique English speakers. Similar to previous experiment, the workers were screened based on being able to identify a word heard in a long sentence.

### Table 1: DDS Dataset - Performance of similarity, no-reference metrics and subjective MOS scores (+0.03) for recordings from the DAPS [31] and VCTK [32] datasets, re-recorded across various environments (DDS dataset) [28]). We see that both *studio2* and *studio3* show close MOS values even though recorded under different environments. We also observe that the similarity metrics rate VCTK recordings as higher quality than the DAPS recordings, even though no-reference metrics and subjective ratings suggest otherwise.

| Type          | DAPS (31) | VCTK (32) |
|---------------|-----------|-----------|
|                | PEQ VISSQL DPAM CDPAM L1 L2 MLP1 SQAPP INDSOS NISQA SQAPP DNSMOS NORESQA MOS | PEQ VISSQL DPAM CDPAM L1 L2 MLP1 SQAPP INDSOS NISQA SQAPP DNSMOS NORESQA MOS |
| Clean         | 3.68 3.45 3.85 9.56 36.48 | 3.44 3.37 3.62 9.97 28.18 |
| Confroom1     | 3.38 3.15 3.80 2.04 22.84 0.17 | 2.81 3.03 3.59 13.24 27.77 |
| Confroom2     | 3.39 3.15 3.80 2.04 22.84 0.17 | 2.81 3.03 3.59 13.24 27.77 |
| Office1       | 1.80 2.42 2.75 0.29 25.86 19.19 | 2.71 3.04 3.42 11.52 26.63 |
| Office2       | 1.68 2.38 2.77 0.32 25.86 19.19 | 2.71 3.04 3.42 11.52 26.63 |
| Studio1       | 2.59 2.35 2.78 0.32 25.90 19.09 | 2.70 3.04 3.42 11.52 26.63 |
| Studio2       | 2.03 2.60 2.70 0.27 24.70 25.18 | 3.48 3.20 3.65 10.91 26.91 |
| Studio3       | 2.03 2.54 2.83 0.29 25.50 30.17 | 3.58 3.18 3.61 10.86 26.83 |
| Waitingroom1  | 2.11 2.55 2.75 0.27 23.78 30.17 | 3.46 3.23 3.55 10.81 26.83 |
| Livingroom1   | 1.56 2.36 2.87 0.30 26.47 12.44 | 2.82 3.14 3.55 11.28 27.28 |

Table 2: Scenario 1 - Performance of similarity and no-reference metrics when recordings from *studio2* and *studio3*, and test recordings from *confroom1* are selected. We see that similarity metrics show different similarities, even though no-reference metrics and subjective ratings (Table 2) suggest the two references are of equal quality.

| Type          | PEQ VISSQL DPAM CDPAM L1 L2 MLP1 SQAPP INDSOS NISQA SQAPP DNSMOS NORESQA |
|---------------|-----------------|-----------------|-----------------|-----------------|
| Clean         | 2.16            | 3.13            | 1.96            |
| L1            | 3.60            | 4.18            | 3.81            |
| L2            | 2.77            | 1.35            | 1.71            |
| L3            | 0.21            | 0.07            | 0.30            |
| L4            | 0.54            | 0.30            |
| L5            | 2.14            | 5.90            |
| Multi-res STFT | 0.14           | 0.07            | 0.11            |
| NISQA         | 4.90            | 4.65            | 3.06            | 4.60            |
| DNSSOS        | 3.64            | 3.55            | 3.02            | 3.55            |
| NORESQA       | 9.57            | 10.53           | 12.88           | 9.44            |
| MOS           | 5.10            | 3.01            | 2.95            | 3.29            |

Table 3: Scenario 2 - Performance of similarity, no-reference metrics and MOS ratings (+0.02) across speech enhancement (SE) models trained on two datasets (DAPS and VCTK), and evaluated on the VCTK evaluation set.
Clean reference | Test recording | Test rec. + high pass filter at 2500Hz

**Figure 4: High-pass filtering** | **Reference, Test, and Test + high-pass filtered recordings shown. PESQ score increases from 2.54 to 2.84, even if visual and perceptual similarity suggests otherwise.**

**Table 4: Scenario 3: Objective measures and MOS ratings (±0.03) across pre-processing stages (Section 2.3) when recordings from DAPS trained SE model are matched to the VCTK trained SE model.**

| Type          | PESQ | VSNR | ESQ | DPAM | L1 | L2 | L1-M1 | L2-M1 | STRFT | SQAPP | NISQA | DNSMOS | NORESQA | MOS |
|---------------|------|------|-----|------|----|----|-------|-------|-------|-------|-------|-------|--------|------|
| VCTK Model    | 2.63 | 4.18 | 2.39 | 2.07 | 0.29 | 2.14 | 0.20 | 2.76 | 4.68 | 3.52 | 10.31 | 10.07 |
| DAPS Model    | 2.50 | 4.12 | 2.61 | 2.42 | 0.19 | 1.95 | 0.15 | 2.50 | 4.08 | 3.63 | 10.31 | 10.07 |
| EQ Match      | 2.52 | 4.68 | 2.71 | 2.22 | 1.95 | 25.99 | 0.13 | 2.96 | 4.89 | 3.61 | 10.32 | 10.11 |
| Breath rem.   | 2.67 | 4.18 | 2.67 | 1.53 | 1.95 | 25.97 | 0.13 | 2.84 | 4.84 | 3.63 | 10.26 | 10.19 |
| Energy norm.  |      |      |      |      |      |      |      |      |      |      |      |        |        |
| Energy norm.  | 2.32 | 4.17 | 2.66 | 0.14 | 1.92 | 24.85 | 0.11 | 2.87 | 4.37 | 3.56 | 10.05 | 10.14 |
| inr200        | 2.29 | 4.17 | 2.66 | 0.14 | 1.92 | 24.85 | 0.14 | 2.84 | 4.34 | 3.53 | 10.07 | 10.10 |
| inr1000       | 2.29 | 4.17 | 2.66 | 0.15 | 1.91 | 24.77 | 0.14 | 2.85 | 4.37 | 3.53 | 10.07 | 10.10 |
| Orig. Phase   | 3.66 | 4.44 | 2.12 | 0.01 | 0.21 | 0.23 | 0.17 | 3.97 | 4.36 | 3.60 | 10.35 | 10.19 |

**3.4. Miscellaneous Observations**

We show several other scenarios where similarity metrics (esp. PESQ) work well, and others that highlight their drawbacks.

**Reversed correlation between PESQ and MOS** Looking at Table 1 specific to PESQ (and other metrics like DPAM, L1, L2, Multi-res. STFT), we observe that the DAPS quality still remains lower than the corresponding VCTK quality for each environment condition, even if subjective listening suggests a reverse trend. This suggests that PESQ (and other metrics) cannot be used to compare absolute qualities across two different datasets (e.g. a PESQ score of 1.55 in confroom1 DAPS v.s. 1.70 VCTK, even when no-reference metrics and MOS ratings suggest otherwise).

**High-pass filtering** Refer to Figure 4. We apply a high pass filter (butterworth filter, order 5) of increasing cutoff frequencies (upto 250Hz) on a test recording, and compare it to a given reference audio recording. We observe that even though the spectrum looks very different, PESQ (and other metrics like DPAM, CDPAM, L1, L2) predict a higher rating with the high-pass filtered test recording as compared to the rating for the clean test recording. Comparing cross-perturbations Refer to Figure 5. We apply different perturbations to the reference recording, and create two test samples: Audio1 contains background white noise, and Audio2 contains compression artifacts. Audio1 sounds closer to the reference according to human judgments. In contrast, similarity metrics find Audio2 closer to the reference (e.g. PESQ score of 2.57 for Audio1 v.s. 3.31 for Audio2, even when subjective listening suggests otherwise) suggesting that these metrics cannot reliably evaluate recordings across perturbations.

**Effect of a noisy reference**: We apply band-limited white noise (between 0 and 3.5kHz) to the reference recording, and observe an increase in PESQ score of up to 0.1 points (Figure 6). Furthermore, when we apply the same noise to both the reference and test recordings (Figure 6), we observe an increase in PESQ of up to 0.2 points. These observations suggest that PESQ can match noise characteristics between signals to output a higher rating.

**Variation with frame-wise alignment** We found PESQ, and other perceptual similarity metrics like VISQOL, DPAM and CDPAM to be quite stable with small mis-alignment in the signal. In contrast, conventional similarity metrics like L1, L2 and multi-resolution STFT show high variance with small mis-alignments.

**4. Conclusions and future work**

In this paper we discussed several scenarios where the conventional formulation of similarity metrics (e.g. PESQ) contrasts from predicting absolute subjective quality, and showed that various no-reference metrics are a good alternative in those settings. Overall, our findings suggest that similarity metrics (like PESQ) are an unreliable proxy for audio quality, and should be used cautiously. In the future, we would like to exhaustively compare the performance of no-reference metrics with full-reference metrics, and see how well they correlate with human perception, especially on out-of-domain tasks.
5. References

[1] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, “Perceptual evaluation of speech quality (PESQ)—a new method for speech quality assessment of telephone networks and codecs,” in ICASSP, vol. 2, 2001, pp. 749–752.

[2] J. G. Beerends, C. Schmidmer, J. Berger, M. Obermann, R. Ullmann, J. Pomny, and M. Keyhl, “Perceptual objective listening quality assessment (POLQA),” the third generation itu-t standard for end-to-end speech quality measurement part I—temporal alignment,” Journal of the AES, no. 6, 2013.

[3] A. Hines, J. Skoglund, A. C. Kokaram, and N. Harte, “ViSQOL: An objective speech quality metric,” EURASIP Journal on Audio, Speech, and Music Processing, vol. 2015, no. 1, 2015.

[4] P. Manocha, A. Finkelstein, R. Zhang, N. J. Bryan, G. J. Mysore, and Z. Jin, “A differentiable perceptual audio metric learned from just noticeable differences,” Interspeech, 2020.

[5] P. Manocha, Z. Jin, R. Zhang, and A. Finkelstein, “CDPAM: Contrastive learning for perceptual audio similarity,” ICASSP 2021.

[6] B. Patton, Y. Agiomygiannakis, M. Terry, K. Wilson, R. A. Saurous, and D. Sculley, “AutoMOS: Learning a non-intrusive assessor of naturalness-of-speech,” arXiv, 2016.

[7] S.-W. Fu, C.-F. Liao, Y. Tsao, and S. Lin, “MetricGAN: Generative adversarial networks based black-box metric scores optimization for speech enhancement,” in ICML, 2019.

[8] M. Yu, C. Zhang, Y. Xu, S. Zhang, and D. Yu, “MetricNet: Improved modeling for non-intrusive speech quality assessment,” Interspeech, 2021.

[9] P. C. Loizou, “Speech quality assessment,” in Multimedia analysis and communications. Springer, 2011.

[10] C. K. Reddy, V. Gopal, and R. Cutler, “DNSMOS: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors,” ICASSP, 2020.

[11] G. Mittag, B. Naderi, A. Chehadi, and S. Möller, “NISQA: A deep CNN-self-attention model for multidimensional speech quality prediction with crowdsourced datasets,” in Interspeech, 2021.

[12] P. Manocha, Z. Jin, and A. Finkelstein, “SQAPP: No-reference speech quality assessment via pairwise preference,” in To Appear ICASSP 2022.

[13] S.-W. Fu, Y. Tsao, H.-T. Hwang, and H.-M. Wang, “Qualitynet: end-to-end non-intrusive speech quality assessment model on blstm,” Interspeech, 2018.

[14] A. H. Andersen, J. M. De Haan, Z.-H. Tan, and J. Jensen, “Nonintrusive speech intelligibility prediction using convolutional neural networks,” IEEE/ACM TASLP, vol. 26, no. 10, pp. 1925–1939, 2018.

[15] C.-C. Lo, S.-W. Fu, W.-C. Huang, X. Wang, J. Yamagishi, Y. Tsao, and H.-M. Wang, “MOSNet: Deep learning based objective assessment for voice conversion,” Interspeech, 2019.

[16] H. Gamper, C. K. Reddy, R. Cutler, I. J. Tashhev, and J. Gehrke, “Intrusive and non-intrusive perceptual speech quality assessment using a convolutional neural network,” in WASPAIA, 2019, pp. 85–89.

[17] Z. Zhang, P. Vyas, X. Dong, and D. S. Williamson, “An end-to-end non-intrusive model for subjective and objective real-world speech assessment using a multi-task framework,” in ICASSP, 2021, pp. 316–320.

[18] J. Su, Z. Jin, and A. Finkelstein, “HiFi-GAN: High-fidelity denoising and dereverberation based on deep speech features in adversarial networks,” Interspeech, 2020.

[19] P. Manocha, B. Xu, and A. Kumar, “NORESQA: A framework for speech quality assessment using non-matching references,” Advances in Neural Information Processing Systems, vol. 34, 2021.

[20] A. Defossez, G. Synnaeve, and Y. Adi, “Real time speech enhancement in the waveform domain,” in Interspeech, 2020.

[21] R. E. Zezario, S.-W. Fu, F. Chen, C.-S. Fuh, H.-M. Wang, and Y. Tsao, “Deep learning-based non-intrusive multi-objective speech assessment model with cross-domain features,” arXiv preprint arXiv:2111.02363, 2021.

[22] J. Su, Z. Jin, and A. Finkelstein, “HiFi-GAN-2: Studio-quality speech enhancement via generative adversarial networks conditioned on acoustic features,” in WASPAIA, Oct. 2021.

[23] M. Cernak and M. Rusko, “An evaluation of synthetic speech using the PESQ measure,” in Proc. European Congress on Acoustics, 2005, pp. 2725–2728.

[24] J. G. Beerends, E. Larsen, N. Iyer, and J. Van Sugts, “Measurement of speech intelligibility based on the pesq approach,” in Proceedings of the Workshop Measurement of Speech and Audio Quality in Networks (MESAQN), Prague, Czech Republic, 2004.

[25] P. Paglierani and D. Petri, “Uncertainty evaluation of speech quality measurement in voip systems,” in 2007 IEEE International Workshop on Advanced Methods for Uncertainty Estimation in Measurement. IEEE, 2007, pp. 104–108.

[26] S.-W. Fu, T.-W. Wang, Y. Tsao, X. Lu, and H. Kawai, “End-to-end waveform utterance enhancement for direct evaluation metrics optimization by fully convolutional neural networks,” IEEE/ACM TASLP, vol. 26, no. 9, pp. 1570–1584, 2018.

[27] Z. Xu, M. Strake, and T. Fingscheidt, “Deep noise suppression maximizing non-differentiable pesq mediated by a non-intrusive pesqnet,” arXiv preprint arXiv:2111.03847, 2021.

[28] H. Li and J. Yamagishi, “DDDS: A new device-degraded speech dataset for speech enhancement,” arXiv preprint arXiv:2109.07931, 2021.

[29] T. Manjunath, “Limitations of perceptual evaluation of speech quality on voip systems,” in 2009 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting. IEEE, 2009, pp. 1–6.

[30] A. Hines, J. Skoglund, A. Kokaram, and N. Harte, “Robustness of speech quality metrics to background noise and network degradations: Comparing ViSQOL, PESQ and POLQA,” in 2013. IEEE, 2013, pp. 3697–3701.

[31] G. J. Mysore, “Can we automatically transform speech recorded in professional production quality speech?—a dataset, insights, and challenges,” IEEE SPS, vol. 22, no. 8, 2014.

[32] C. Valentini-Boitinhao et al., “Noisy speech database for training speech enhancement algorithms and TTS models,” 2017.

[33] E. Cooper, W.-C. Huang, T. Toda, and J. Yamagishi, “Generalization ability of mos prediction networks,” arXiv preprint arXiv:2110.02635, 2021.

[34] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, “A short-time objective intelligibility measure for time-frequency weighted noisy speech,” in ICASSP. IEEE, 2010, pp. 4214–4217.

[35] F. G. Germain, G. J. Mysore, and T. Fujikawa, “Equalization matching of speech recordings in real-world environments,” in ICASSP. IEEE, 2016, pp. 609–613.

[36] A. Pandey and D. Wang, “On cross-corpus generalization of deep learning based speech enhancement,” IEEE/ACM TASLP, vol. 28, pp. 2489–2499, 2020.