A NO-REFERENCE VIDEO QUALITY PREDICTOR FOR COMPRESSION AND SCALING ARTIFACTS

Deepti Ghadiyaram

The University of Texas at Austin

Chao Chen, Sasi Inguva, and Anil Kokaram

Google Inc.

ABSTRACT

No-Reference (NR) video quality assessment (VQA) models are gaining popularity as they offer scope for broader applicability to user-uploaded video-centric services such as YouTube and Facebook, where the pristine references are unavailable. However, there are few, well-performing NR-VQA models owing to the difficulty of the problem. We propose a novel NR video quality predictor that solely relies on the 'quality-aware' natural statistical models in the space-time domain. The proposed quality predictor called **S**elf-reference based **LE**arning-free **E**valuator of **Q**uality (SLEEQ) consists of three components: feature extraction in the spatial and temporal domains, motion-based feature fusion, and spatial-temporal feature pooling to derive a single quality score for a given video. SLEEQ achieves higher than 0.9 correlation with the subjective video quality scores on tested public databases and thus outperforms the existing NR VQA models.

***Index Terms*—** Perceptual video quality, objective quality assessment, H.264 compression, scaling artifacts.

1. INTRODUCTION

With the pervasiveness of visual media, measuring and improving the quality of images and videos is receiving considerable attention. Digital videos often contain visual distortions that are introduced during capture, storage, or transmission. These distortions often detract from a viewer's quality of experience (QoE). The visual quality of the consumed video content *as perceived by human observers* is termed as 'perceptual quality.' Understanding and *objectively* determining human observers' *subjective* perception of video quality is of great importance to camera designers, video streaming services such as Netflix and YouTube, network and content providers and many more visual-media-driven services. Thus, the development of automatic objective methods that accurately quantify the impact of visual distortions on perception has greatly accelerated.

The goal of no-reference (NR) objective video quality assessment (VQA) algorithms is the following: given an input video and no additional information, accurately predict its perceptual quality. Though the full-reference (FR) quality predictors such as SSIM [1], VQM [2], and PSNR are fast and accurate, they require a pristine reference video signal with respect to which the quality of the distorted video is assessed. In the context of cloud video transcoding systems such as YouTube, these FR metrics consider the original uploaded video as a reference video. The videos transcoded at different bitrates and resolutions (in order to tailor to the plethora of consumer display devices) are considered as distorted videos and their quality is measured with respect to the uploaded video. Considering a user-uploaded video to be pristine, i.e., void of any naturally occurring distortions is a highly incorrect assumption, since, a large fraction of the uploaded videos are certainly compressed prior to uploading on YouTube. Thus, the resulting signal fidelity measured against the uploaded video does not reflect the true perceived quality of the transcoded videos and could lead to incorrect transcoding choices. On the other hand, having an accurate and fast NR-VQA model that could detect the quality of the uploaded and transcoded videos could assist in designing 'quality-aware' encoding processes such as perceptually choosing the appropriate encoding bitrates. Such strategies could in turn help ensure that end users enjoy satisfactory levels of QoE. NR VQA models can also aid in identifying and culling low quality videos stored on digital devices, designing and verifying suitable video quality-enhancement processes, and so on.

Nowadays, video streaming services are typically leveraging HTTP-based adaptive streaming protocols such as Dynamic Adaptive Streaming over HTTP (DASH) and HTTP Live Streaming (HLS). Thus, packet losses and bit errors are no longer the main sources of visual impairments in the transmitted videos. However, videos need to be downsampled and compressed at different bitrates to adapt to client-side network bandwidth and display devices. This subsequently introduces *compression* and *scaling* artifacts in the video delivered to viewers. Thus, we focus exclusively on designing a no-reference quality predictor for videos afflicted with H.264 compression and scaling artifacts.

Related Work: Video quality assessment is an active area of research, and several VQA algorithms have been designed recently to *separately* address compression and upscaling artifacts [3, 4, 5, 6]. V-BLIINDS [7], V-CORNIA [8], and VIIDEO [9] are some of the recent top-performing distortion-agnostic NR VQA algorithms. To the best of our knowledge,

Netflix's recently introduced full-reference model VMAF [10] is the only VQA algorithm designed for videos where compression and scaling artifacts occur simultaneously. Due to space constraints, we refer the reader to [9] for a thorough review of the state-of-the-art VQA algorithms.

Given that video and subjective data collection is a laborious and a time-consuming process, most existing video datasets are typically smaller in size (only 40 H.264 compressed videos in LIVE VQA Database [11] and 70 publicly-available distorted videos in Netflix's database [12]). Training-based VQA models could thereby lead to overfitting to these small-sized datasets and might not generalize well over real-world videos uploaded in YouTube. Thus, designing an accurate and a training-free model that would generalize across databases and video contents is highly desirable.

Unlike most existing VQA algorithms, our proposed model called **S**elf-reference based **LE**arning-free **E**valuator of **Q**uality (SLEEQ) is 'opinion-unaware' and is thus 'completely blind.' Specifically, SLEEQ uses natural scene statistics (NSS) based models that characterize natural videos in spatial and temporal domains to extract a single quality-informative feature, thus not requiring any form of training on video datasets. These virtues combined with SLEEQ's accurate quality predictions offer a significant advantage over existing FR and learning-based NR-VQA models.

We now proceed to discuss the details of our NR VQA model SLEEQ in Sec. 2 and report its performance and other FR/NR VQA models on public VQA databases in Sec. 3.

2. DETAILS OF THE PROPOSED ALGORITHM

2.1. Natural Video Statistics

Current efficient NR Image Quality Assessment (IQA) algorithms [13, 14] use natural scene statistics (NSS) based models to capture the 'statistical naturalness' of images that are not distorted [15]. NSS models rely on the fundamental observation that suitably normalized coefficients of good quality real-world photographic images follow a Gaussian distribution [16], which the distorted images deviate from. This deviation from statistical regularity holds information about the underlying distortions afflicting these images [13]. More recently, similar statistical regularities and irregularities were observed to be exhibited in the temporal domain by the *frame differences* of the consecutive frames in videos [17, 7, 9]. Therefore, in our work, we effectively quantify both spatial and temporal scene statistics to extract quality-aware features and combine them to be able to make predictions regarding the perceptual quality of videos.

Specifically, given an input I of size $M \times N$, which could either be the luminance component[1] of a frame (f_n) or difference between the luminance components of consecutive frames ($d_n = f_{n+1} - f_n$), where n is a frame index, a divisive normalization operation can be applied on I, which is defined as follows:

$$\mathrm{N}(i,j) = \frac{I(i,j) - \mu(i,j)}{\sigma(i,j) + 1}, \tag{1}$$

where

$$\mu(i,j) = \sum_{k=-K}^{K} \sum_{\ell=-L}^{L} \mathrm{w}_{k,\ell} I(i-k, j-\ell) \tag{2}$$

and

$$\sigma(i,j) = \sqrt{\sum_{k=-K}^{K} \sum_{\ell=-L}^{L} \mathrm{w}_{k,\ell} \left[I(i-k, j-\ell) - \mu(i-k, j-\ell)\right]^2} \tag{3}$$

where $i \in \{1, 2, \cdots, M\}, j \in \{1, 2, \cdots, N\}$ are spatial indices and $\mathbf{w} = \{w_{k,l} | k = -K, ..., K, l = -L, ...L\}$ is a 2D circularly-symmetric Gaussian weighting function (with $K = L = 3$, and thus standard deviation $= 1.16$). It was found that a generalized Gaussian distribution (GGD) effectively models the statistics of these mean subtracted and contrast normalized (MSCN) coefficients [13]. A GGD with zero mean is given by:

$$f(x; \alpha, s^2) = \frac{\alpha}{2\beta\Gamma(1/\alpha)} exp\left(-\left(\frac{|x|}{\beta}\right)^{\alpha}\right), \tag{4}$$

where

$$\beta = s\sqrt{\frac{\Gamma(1/\alpha)}{\Gamma(3/\alpha)}} \tag{5}$$

and $\Gamma(.)$ is the gamma function:

$$\Gamma(a) = \int_0^{\infty} t^{a-1} e^{-t} dt \quad a > 0. \tag{6}$$

The two parameters of the GGD distribution: shape (α) and standard deviation (s) can be estimated using an efficient moment-matching based approach [18].

In our algorithm, we partition I (both f_n and d_n) into non-overlapping patches and compute GGD-shape (α) values for all the patches of I across alternate frames of a video, which are later average pooled (see complete details in Sec. 2.4).

2.2. Content Dependency problem with the NSS-features

Though Gaussianity is exhibited by most pristine video frames (and frame differences), we empirically found that the deviation from Gaussianity exhibited by distorted videos is *not* consistent across videos with different contents. To illustrate this behavior, in Fig. 1 (Left), along the X-axis, we plot the average of the GGD shape α computed across all the patches of all the frames of 50 videos[2] from LIVE VQA Database [11]. Along the Y-axis, we plot the ground truth quality scores, i.e., the Difference Mean Opinion Scores (DMOS) associated with each video. A lower value of DMOS indicates better quality of the video. Each color and marker

[1] The chroma components of a video frame are not considered in our work.

[2] We considered only 40 videos afflicted with H.264 compression and their pristine counterparts (10 videos), thus making a total of 50 videos.

in this plot correspond to a different video content (pristine video and its corresponding distorted variants).

From this plot, it can be observed that these video features mostly take similar values ($>= 2.0$) for all ten pristine videos (with DMOS = 0) as has already been observed in the case of pristine images in [13, 14]. However, in the case of distorted videos, these features take different values for different video contents. Despite the consistent monotonicity between the video features and DMOS for *a given video content*, these feature values are not consistent across different contents that have similar DMOS scores. For instance, for videos with DMOS values between $30 - 40$, the feature values varied between $1.6-2.6$, which is highly undesirable. It is probable that the interaction of the spatial and temporal content and compression artifacts is resulting in such inconsistent feature values for different video contents. Similar content-dependency concerns were reported in [7] in the DCT domain as well, where the absolute parameter values were found to be less reliable for quality prediction.

2.3. 'Self-Reference'-based Feature Extraction

As a way to tackle this content-dependency issue, we propose the following solution. Given a frame (f_n) and a frame difference image (d_n), we apply a low-pass filter to them and construct their blur partners f'_n and d'_n. We then extract the features described in Sec. 2.1 from the patches of the original frames (and frame differences) and their blur variants, and compute an absolute difference of these features. We then take an average of these values over all the video frames and use this as our video quality feature (more details in Sec. 2.4).

Our motivation to construct a blur variant of a video and use the given video itself as its reference (hence called self-referencing) was the following: though the deviation in Gaussianity is content-dependent (as observed in Fig. 1 (Left)), we hypothesized that the *difference* in the deviation from Gaussianity between the given video and its blur variant might be consistent across different video contents. This *self-referencing* technique thus could effectively capture the relative inter-dependency between the NSS features of the given video and its blur variant and could potentially aid in reducing the content-dependency issue.

Figure 1 (Right) illustrates the benefit of adopting self-referencing technique. Here, along the X-axis are the self-referenced video quality features Q (defined in Sec. 2.4) of the same 50 videos of the LIVE-VQA Database used in Fig. 1 (Left) and along the Y-axis are the corresponding DMOS values. We can observe that the severe content dependency observed in Fig. 1 (Left) has reduced to a great extent in Fig. 1 (Right). It can also be observed that the strong monotonicity between the quality features and DMOS values persists independent of the video content.

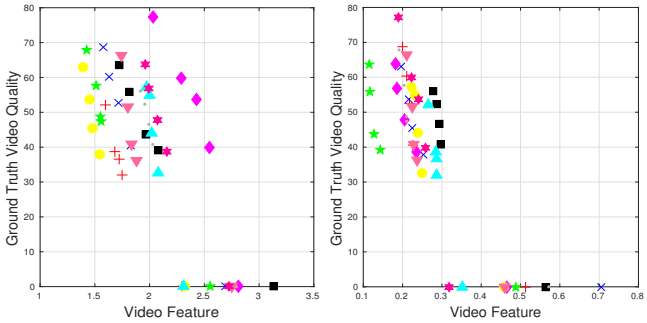

Fig. 1. Along the X-axis of these scatter plots are the (Left) aggregated (GGD) shape features and (Right) self-referenced video quality features. Along the Y-axis are the ground truth quality scores of 50 videos in the LIVE VQA Database [11]

2.4. Feature Extraction, Fusion, and Pooling

Self-referencing technique is at the core of SLEEQ and to the best of our knowledge has not been used in the quality assessment literature. We now describe all the components of our algorithm. As mentioned earlier, we consider the luminance component of every alternate frame (f_n) and compute a frame difference image (d_n), where n is the frame index. We apply a low-pass filter (a 2-D Gaussian smoothing kernel with a standard deviation B_σ) to them and obtain their blurred variants f'_n and d'_n. We divide these four images into patches of size $P \times P$ and apply the divisive normalization operator (as defined in 1) on patches of all four images independently. We extract a total of 4 GGD shape features from each patch: α_s^P, $\alpha_s'^P$ are the spatial features extracted from f_n and f'_n, and α_t^P, $\alpha_t'^P$ are the temporal features extracted from d_n and d'_n respectively. Further, as discussed in Sec. 2.3, we compute an absolute difference of the spatial and temporal GGD-parameters as follows:

$$\Delta\alpha_s^P = |\alpha_s'^P - \alpha_s^P|, \tag{7}$$

$$\Delta\alpha_t^P = |\alpha_t'^P - \alpha_t^P| \tag{8}$$

Weighting features based on the motion: Perceptual quality of a video depends on both spatial and temporal content and their interactions with the distortions afflicting them. As a way to combine the spatial and temporal quality-aware features (defined in (7) - (8)), we adopted the following strategy: In parts of the video where there is little motion, we assigned more importance to spatial features ($\Delta\alpha_s^P$). In such scenarios, since the frame differences do not contain sufficient information to capture structural regularity (or irregularity) due to distortions, we hypothesized that the spatial quality would dominate a viewer's perception. Conversely, in parts of the video where there is significant motion, the presence of visual distortions could further degrade its perceived quality. Therefore, we assigned more importance to temporal features ($\Delta\alpha_t^P$) as temporal distortions could dominate a viewer's perceived quality. This weighing scheme is formulated as below.

$$Q_p = (1 - m_p) \cdot \Delta\alpha_s^P + m_p \cdot \Delta\alpha_t^P, \tag{9}$$

where m_p is simply the normalized average frame difference in a given patch P, and $\Delta\alpha_s$ and $\Delta\alpha_t$ are the spatial and

temporal features computed on the patch P.

Spatial-temporal pooling: Since humans appear to more heavily weight their judgments of image quality from the sharp image regions [19], more salient quality measurements can be made from sharp patches. We thus use a simple technique to preferentially select from amongst a collection of patches per frame, those that are richest in spatial information. Towards this end, we compute an average standard deviation (defined in (3)) of every patch in f_n and f'_n, denoted as $\bar{\sigma_p}$ and $\bar{\sigma}'_p$ respectively. We then compute a difference of the average standard deviation per patch given by,

$$\Delta\bar{\sigma}_p = |\bar{\sigma}'_p - \bar{\sigma_p}|. \tag{10}$$

We exclude patches whose $\Delta\bar{\sigma}_p$ value is below n-th percentile of all the $\Delta\bar{\sigma}_p$ values of a given video. Following this, we average-pool the motion-weighted features (Q_p) from the filtered patches of a frame and across all the frames in a given video and compute a single final quality score (Q).

3. EXPERIMENTAL RESULTS

This section reports the performance of our proposed self-referenced algorithm SLEEQ and a few others on two different datasets - the 40 H.264 compressed videos (and the 10 pristine videos) all of resolution 768×432 in the LIVE VQA Database [11] and the 79 publicly-available videos all of resolution 1920×1080 from the Netflix dataset [10]. The publicly available source codes of different NR/FR VQA models were used to extract features and train models on the two databases. VMAF [20] and V-BLIINDS [7] require training, thus, each dataset for every experiment was split into content-independent, non-overlapping training and test datasets: 80% of the content was used for training and the remaining 20% was used for testing. To mitigate any bias due to the division of data, all possible combinations of content splits were considered. VMAF and V-BLIINDS were trained from scratch on 80% of the data and tested on the remaining non-overlapping 20% test data. For VIIDEO [9] and SLEEQ, no learning is required, yet, for a fairer comparison with the other learning-based models, we only report the median correlations on 20% of the test data. Spearman's rank ordered correlation coefficient (SROCC) and Pearson's correlation coefficient (PLCC) between the predicted and the ground truth quality scores are reported. A higher value of each of these metrics indicates better performance in terms of correlation with human opinions. We also report the average running time (in seconds) for extracting quality features designed in each VQA algorithm computed over 5 random chosen videos[3] from both the databases. Scores predicted from our model and that of VIIDEO were passed through a 5 parameter non-linear logistic mapping function [21, 22] before computing PLCC.

[3]The chosen videos from the LIVE VQA Database were composed of 450 frames, while those from the Netflix database had 150 frames.

Table 1. Performance on the 50 H.264 compressed videos of the LIVE VQA Database [11]

VQA Type	VQA Model	Avg. Run Time	SROCC	PLCC
NR	SLEEQ	25 sec.	0.90	0.96
NR	VIIDEO [9]	93 sec.	0.76	0.89
NR	V-BLIINDS [7]	307 sec.	0.79	0.88
FR	VMAF [20]	25 sec.	0.96	0.97

Table 2. Performance on the 79 videos of the Netflix dataset [10].

VQA Type	VQA Model	Avg. Run Time	SROCC	PLCC
NR	SLEEQ	96 sec.	0.93	0.91
NR	VIIDEO [9]	500 sec.	-0.49	-0.6
NR	V-BLIINDS [7]	1545 sec.	0.92	0.90
FR	VMAF [20]	95 sec.	0.95	0.95

The two parameters in the proposed algorithm, (a) n-th percentile threshold for $\Delta\bar{\sigma}_p$ and (b) the blur kernel's standard deviation B_σ, vary with a test video's resolution. In particular, for videos of larger resolution, higher values of these parameters improved the performance of our model. We found via cross-validation that $B_\sigma \in [1, 3]$ and $n \in [5, 10]$ for videos in LIVE VQA Database and $B_\sigma \in [7, 11]$ and $n = [35, 40]$ for videos in Netflix's database yielded a high overall performance. In our experiments, we used a patch size of $P = 72$ for both the datasets, $n = 5$, $B_\sigma = 1.16$ for the LIVE VQA Database and $n = 35$, $B_\sigma = 11$ for the Netflix Database.

From the results reported in Tables 1 and 2, we can observe that SLEEQ outperforms the state-of-the-art NR VQA models on both the databases and competes very well with the full-reference model VMAF. It can also be observed that an unoptimized MATLAB implementation of SLEEQ with no parallelization is much faster than the other NR VQA models, and compare very well with the highly optimized implementation of VMAF. VIIDEO captures only temporal statistics of natural videos and thus its performance suffers on Netflix database which contain videos with rich spatial content and mixtures of compression and scaling artifacts, in addition to object motion and ego-motion. In summary, SLEEQ is much faster than the existing NR-VQA models, training-free, and has superior prediction performance.

4. CONCLUSION AND FUTURE WORK

We proposed a natural scene statistics-based VQA method called SLEEQ for accurately predicting the perceived quality of videos afflicted with H.264 compression and scaling artifacts. Every component of the proposed algorithm is computationally very simple and parallelizable and thus deployable in real-world applications. While VMAF [20] also tackles compression and scaling artifacts, it is a full-reference and a learning-based model. SLEEQ, on the other hand, is no-reference and *training-free* and yet serves as an excellent indicator of video quality. We plan to optimize our feature extraction process, evaluate our model on other datasets, and extend our model to other distortions, primarily real-world distortions that typically occur in the mobile videos [23, 24] in the future.

5. REFERENCES

[1] Z. Wang, A.C Bovik, H.R Sheikh, and E.P Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Trans. on Image Proc.*, vol. 13, no. 4, pp. 600–612, 2004.

[2] F. Xiao, "DCT-based video quality evaluation," *Final Project for EE392J Stanford Univ.*, 2000.

[3] X. Lin, H. Ma, L. Luo, and Y. Chen, "No-reference video quality assessment in the compressed domain," *IEEE Trans. on Consumer Electronics*, vol. 58, no. 2, pp. 505–512, 2012.

[4] C. Keimel, T. Oelbaum, and K. Diepold, "No-reference video quality evaluation for high-definition video," in *IEEE Int. Confernce on Acoustics, Speech and Signal Proc.*, 2009, pp. 1145–1148.

[5] K. Zhu, K. Hirakawa, V. Asari, and D. Saupe, "A no-reference video quality assessment based on laplacian pyramids," in *IEEE Int. Conf. on Image Proc.*, 2013, pp. 49–53.

[6] H. Yeganeh, M. Rostami, and Z. Wang, "Objective quality assessment of interpolated natural images," *IEEE Trans. on Image Proc.*, vol. 24, no. 11, pp. 4651–4663, 2015.

[7] M. A. Saad, A. C. Bovik, and C. Charrier, "Blind prediction of natural video quality," *IEEE Trans. on Image Proc.*, vol. 23, no. 3, pp. 1352–1365, March 2014.

[8] J. Xu, P. Ye, Y. Liu, and D. Doermann, "No-reference video quality assessment via feature learning," in *IEEE Int. Conf. on Image Proc.*, 2014, pp. 491–495.

[9] A. Mittal, M. A. Saad, and A. C. Bovik, "A completely blind video integrity oracle," *IEEE Trans. on Image Proc.*, vol. 25, no. 1, pp. 289–300, Jan 2016.

[10] "VMAF-Video Multi-Method Assessment Fusion," [Online] Available, `https://github.com/Netflix/vmaf/`.

[11] K. Seshadrinathan, R. Soundararajan, A. C. Bovik, and L. K. Cormack, "Study of subjective and objective quality assessment of video," *IEEE Trans. on Image Proc.*, vol. 19, no. 6, pp. 1427–1441, June 2010.

[12] A. Aaron, Z. Li, M. Manohara, J. Y. Lin, E. C. H. Wu, and C. C. J. Kuo, "Challenges in cloud based ingest and encoding for high quality streaming media," in *IEEE Int. Conf. on Image Proc.*, Sept 2015, pp. 1732–1736.

[13] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Trans. on Image Proc.*, vol. 21, no. 12, pp. 4695–4708, Dec 2012.

[14] D. Ghadiyaram and A.C. Bovik, "Perceptual quality prediction on authentically distorted images using a bag of features approach," *Journal of Vision*, https://arxiv.org/abs/1609.04757.

[15] A. C. Bovik, "Automatic prediction of perceptual image and video quality," *Proceedings of the IEEE*, vol. 101, no. 9, pp. 2008–2024, Sept 2013.

[16] D. L. Ruderman, "The statistics of natural images," *Network: Computation in Neural Systems*, vol. 5, no. 4, pp. 517–548, 1994.

[17] R. Soundararajan and A. C. Bovik, "Video quality assessment by reduced reference spatio-temporal entropic differencing," *IEEE Trans. on Circuits and System for Video Tech.*, vol. 23, no. 4, pp. 684–694, 2013.

[18] K. Sharifi and A. Leon-Garcia, "Estimation of shape parameter for generalized gaussian distributions in subband decompositions of video," *IEEE Trans. on Circuits and Systems for Video Tech.*, vol. 5, no. 1, pp. 52–56, Feb 1995.

[19] R. Hassen, Z. Wang, and M. Salama, "No-reference image sharpness assessment based on local phase coherence measurement," pp. 2434–2437, 2010.

[20] "Toward a practical perceptual video quality metric," Accessed: 2016-06-06, [Online] Available: `http://techblog.netflix.com/2016/06/toward-practical-perceptual-video.html/`.

[21] "ITU Objective perceptual assessment of video quality: Full reference television. 2004," [Online] Available: `https://www.itu.int/ITU-T/studygroups/com09/docs/tutorialopavc.pdf`.

[22] H. R. Sheikh, M. F. Sabir, and A. C. Bovik, "A statistical evaluation of recent full reference image quality assessment algorithms," *IEEE Trans. on Image Proc.*, vol. 15, no. 11, pp. 3440–3451, Nov 2006.

[23] M. Nuutinen, T. Virtanen, M. Vaahteranoksa, T. Vuori, P. Oittinen, and J. Hkkinen, "CVD2014;A Database for Evaluating No-Reference Video Quality Assessment Algorithms," July 2016, vol. 25, pp. 3073–3086.

[24] D. Ghadiyaram, J. Pan A.C. Bovik, A. K. Moorthy, P. Panda, and K. C. Yang, "Subjective and objective quality assessment of mobile videos with in-capture distortions," *IEEE Int. Conf. on Acoustics, Speech and Signal Proc.*, 2017, (in print).