INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

An Objective Evaluation Methodology for Blind Bandwidth Extension

Stéphane Villette, Sen Li, Pravin Ramadas, Daniel J. Sinder

Qualcomm Technologies, Inc., 5775 Morehouse Drive, San Diego, CA 92121-1714, United States
{svillett, senl, pramadas, dsinder}@qti.qualcomm.com

Abstract

In this paper we introduce an objective evaluation methodology for Blind Bandwidth Extension (BBE) algorithms. The methodology combines an objective method, POLQA, with a bandwidth requirement, based on a frequency mask. We compare its results to subjective test data, and show that it gives consistent results across several bandwidth extension algorithms. Additionally, we show that our latest BBE algorithm achieves quality similar to AMR-WB at 8.85 kbps, using both subjective and objective evaluation methods.

Index Terms: blind bandwidth extension, artificial bandwidth extension, speech coding, objective quality evaluation, subjective quality evaluation, POLQA

1. Introduction

Until a few years ago, the quality of voice telecommunications has been limited by design choices made over 100 years ago, which resulted in a 8 kHz sampling rate being used and in a practical frequency range of 300-3400Hz. This so called Narrowband (NB) frequency range severely limited speech quality. Recently, the industry has started to move to "HD voice" and "Ultra HD voice", i.e. the use of wideband (WB) or super-wideband (SWB) coders, respectively, which use a sampling rate of 16kHz or 32kHz and correspond to a frequency range of 50-7000Hz or 50-14000Hz respectively [1] [2].

However, these deployments are not ubiquitous. A whole new infrastructure is needed to support these WB and SWB coders, at a substantial cost. While progress is being made, deployment is still limited, and it will likely take years before complete coverage is achieved. Until then, a significant proportion of calls will still use legacy narrowband. Further, it is likely that landline upgrades to WB or SWB will take even longer, meaning that even when the mobile networks have fully migrated to higher bandwidths, calls from landlines will still be narrowband.

2. BBE and Objectives

Blind Bandwidth Extension (BBE) technology aims at solving this problem, by transforming NB speech into WB or SWB speech. In this paper we will focus on the WB case only for simplicity. Typically using some form of either spectral folding or statistical modelling, the 4-8 kHz part of a speech signal is predicted from the 0-4 kHz part, to generate a signal having the general characteristics of wideband speech [3][4]. While perfect prediction cannot be expected, reasonably good quality speech can be obtained.

There are two ways to view the objectives of BBE. It can either be seen as a way to improve NB, or as a way to make NB closer to WB. While these may seem like very similar objectives, in practice they are quite different, and apply to different scenarios. The first case is that of a network that is currently NB only, while the second case is encountered when a network has a mix of NB and WB calls. Both of these scenarios are encountered across mobile phone networks. As networks move towards deploying more HD voice codecs, the second scenario will become more common. The user will likely experience a mix of wideband and narrowband calls, or possibly even experience both bandwidths during the same call. The lack of uniformity of experience will be a problem, as some calls will appear muffled or lower quality, which in turn will lead to user dissatisfaction.

3. BBE Quality Evaluation

3.1. Challenge: Bandwidth vs quality

BBE evaluation has proven to be a challenging task. There is currently no well-established way of evaluating BBE performance. The main issue with BBE quality evaluation is that BBE algorithms are not perfect, and the process of predicting a high-band tends to introduce artifacts. Therefore, for a given BBE algorithm, there is a trade-off between bandwidth of the signal and overall noisiness of the BBE extended speech. This bandwidth/quality trade-off can be controlled easily, e.g. by attenuating the overall high-band energy. A given algorithm can be tuned to offer very little high-band energy, and very few artifacts. Alternatively, it can be tuned to offer levels of high-band equal to that of WB speech, at the cost of more artifacts.

This can lead to confusion during comparative evaluations, where listeners might prefer an algorithm because it shows fewer artifacts when this is in fact due to it having less high-band energy, rather than being intrinsically a better algorithm. Therefore, it is important that different BBE algorithms are compared at the same operating point. This is illustrated in Figure 1.

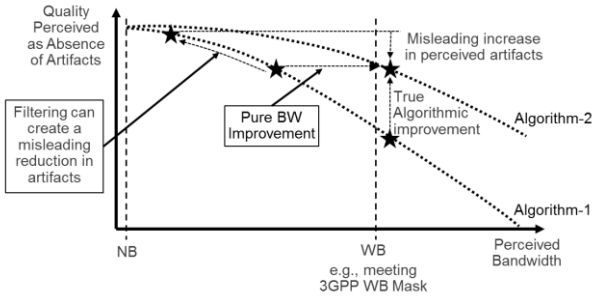

Figure 1: Bandwidth vs Absence of artifacts trade-off

In Figure 1, two BBE algorithms are represented. Algorithm-2 is clearly better than Algorithm-1. This is easily seen when

http://dx.doi.org/10.21437/Interspeech.2016-1595

fixing one dimension, either bandwidth or quality: Algorithm-2 is superior in the other dimension. The problem occurs when comparing Algorithm-1 at low bandwidth (the operating point furthest to the left), to Algorithm-2 at high bandwidth (the operating point furthest to the right). In this situation, Algorithm-1 has fewer artifacts than Algorithm-2, even though the algorithm itself is not as good, only the operating points are different. This shows the necessity of considering both dimensions when comparing BBE algorithms.
Additionally, as bandwidth is reduced, all BBE algorithms converge to the input narrowband signal, and are indistinguishable. Therefore, for maximum resolution, it is best to evaluate BBE algorithms at a high bandwidth, even if it might not be the bandwidth at which the algorithm is intended to be used for deployment.

3.2. Defining bandwidth

Bandwidth of a speech processing systems, such as a vocoder, is usually estimated by comparing its overall frequency response to a frequency mask, *e.g.*, as in [5]. But this is not possible for BBE technologies, as the high-band of the output is independent of the high-band of the input (which is zero in the input narrowband signal). Therefore, the frequency response for BBE is not defined. This problem can be resolved by defining a reference wideband input, to be used for this calculation. The speech material defined in the ITU-T P.501 standard [6] is a good choice since it is broadly used across the wireless industry for testing compliance for voice services. This material is both freely available and already included in many vendors' test equipment.

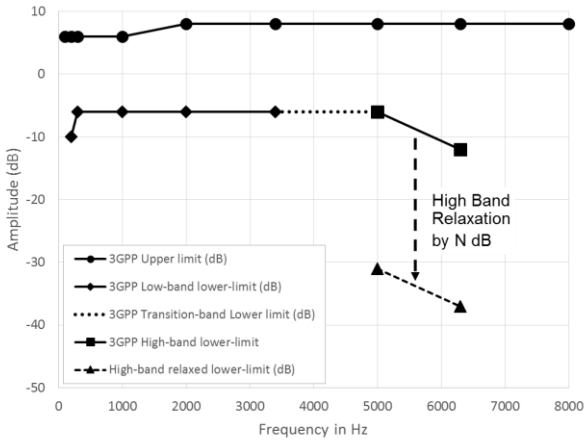

Figure 2: Frequency mask for bandwidth estimation

With regards to the frequency mask, the 3GPP WB Rx mask defined in [5] is also a widely used standard implemented in test equipment, making it a good mask to choose for use with WB BBE. This likewise ensures that the bandwidth of the BBE output is similar to that of a coded, wideband output meeting the same mask. However, to allow for a different operating point at lower bandwidth, a series of masks can be defined as modifications to the 3GPP WB Rx mask wherein its lower limit is relaxed by N dB in the high-band. This is illustrated in Figure 2. Note that the 3.3–5kHz transition-band has been left undefined, to allow for classic frequency extension techniques such as spectral folding, which can lead to a frequency dip around 4 kHz without adversely affecting speech quality.

3.3. Subjective and objective evaluation methods

The most commonly used techniques for subjective quality evaluation of vocoders are the ITU-T P.800 DCR (Degradation Category Rating) and ACR (Absolute Category Rating) tests [7]. Both are suitable for BBE evaluation, the main difference being that DCR measures degradation from the WB reference input, whereas ACR does not present a reference. Interestingly, these two cases match the two deployment scenarios described above, with DCR corresponding to the NB/WB mixed network case, and ACR to the NB-only case.
However, subjective tests are costly and time-demanding. An increasingly popular alternative is to use objective evaluation methods, in particular ITU-T P.863, also known as POLQA [8]. While it is not perfect, POLQA claims to handle a wide range of input degradations, and when used appropriately, can give a good indication of subjective speech quality [9]. Additionally, it is already widely used in the industry for speech quality evaluation, often with ITU-T P.501 source material [6]. For BBE, the source material should be transcoded by an appropriate narrowband vocoder. If cellular wireless transmission is under consideration, this most commonly means the 3GPP AMR codec operating at 12.2 kbps [10], as this is the narrowband speech codec used in the vast majority of today's mobile communication networks.

3.4. Proposed BBE objective evaluation methodology

In summary, we propose the following objective evaluation methodology for BBE.

- Bandwidth requirement:
  - Measure bandwidth by testing the response to verify whether it passes a frequency mask derived from the 3GPP WB Rx mask, as per Figure 2, and using ITU-T P.501 British English speech material as the input.
  - We recommend using N=0 dB (i.e., no relaxation of the mask) as the operating point. Looser requirements can be set.
- Quality requirement:
  - Measure quality using POLQA with P.501 British English coded by AMR at 12.2kbps.
  - A good quality reference is the POLQA score of the input NB signal, up-sampled to 16 kHz.

Note that POLQA has a number of options and versions. In this paper, we are using POLQA v2.4, in High-Accuracy mode, and a WB reference. Other options change the absolute POLQA scores, but generally have little impact on the relative scores, and do not change the overall conclusions of this paper.

4. BBE algorithm evaluation

4.1. Algorithms used

To illustrate the various evaluation techniques, we have evaluated four BBE algorithms according to the above proposed methodologies.
The four BBE algorithms evaluated are:

- **BBE1** is a simple noise addition algorithm in which the high-band is random noise scaled by the energy of the low-band on a 20ms basis. This is included for illustrative purposes, as a simplistic form of bandwidth extension. Its subjective quality is poor, and it is not suitable for field use.

- **BBE2** and **BBE3** are two proprietary blind bandwidth extension technologies that are commercially available in Qualcomm products.
- **BBE4** is a newer algorithm, currently in R&D stages.

The input to these BBE algorithms is narrowband PCM transcoded by AMR at 12.2 kbps, which is the most commonly used speech coder on mobile networks, and is also equivalent to EFR. [10]

4.2. Objective performance

To illustrate the importance of the two-factor approach presented in section 3.4, we plotted the POLQA scores for these BBE algorithms versus their bandwidth. Each algorithm is designed to meet the 3GPP WB Rx mask, and the mask is progressively relaxed from 0 to 25 dB attenuation. As the mask is relaxed, the algorithm output is filtered correspondingly. The scores for AMR NB at 12.2 kbps and AMR-WB at 8.85 kbps are shown as references. The results are shown in Figure 3.

As expected, POLQA scores drop as the bandwidth of the signal gets closer to WB. In effect, POLQA heavily penalizes over-predicting high band energy, and reducing the amount of predicted high-band energy overall helps to improve the POLQA score, even as the subjectively perceived bandwidth decreases. As the mask is relaxed further, the scores flatten out. This is expected, as the lower limit of the mask is a minimum requirement for high-band energy, but of course the signal does not have to follow the mask attenuation.

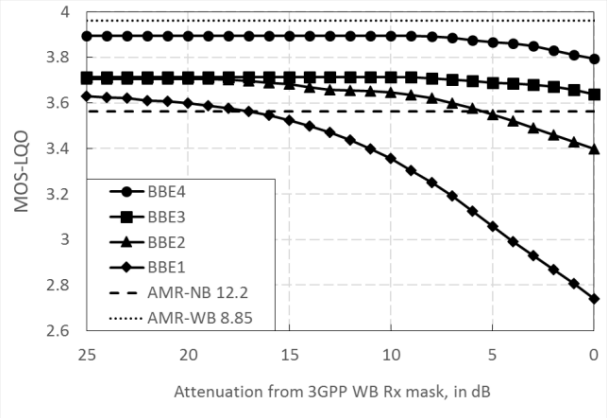

Figure 3: POLQA MOS-LQO vs Bandwidth

There are several interesting points shown in Figure 3. Firstly, it can be seen that BBE technology can provide a significant objective quality advantage over narrowband, and approach the quality of AMR-WB at 8.85 kbps. Indeed BBE4 scores up to 0.35 MOS-LQO higher than the narrowband reference.
Secondly, even BBE1, a very basic BBE algorithm with poor audio quality, can outperform the original narrowband, up to approximately the 18 dB attenuation point. This clearly indicates that quality (as measured by POLQA) is not a reliable indicator by itself, and must be considered in conjunction with bandwidth.
Finally, even though BBE2 and BBE3 achieve similar POLQA score at high attenuation, BBE3 is able to maintain that performance much better than BBE2 as bandwidth increases. Therefore, for reliable discrimination between BBE algorithms, the most interesting measurements are the attenuation at the cross-over point with the narrowband reference, and the POLQA score at 3GPP mask level. (I.e. the 0 dB point on the curve).

4.3. Subjective performance

The subjective performance of the various BBE algorithms presented here was evaluated using the ITU-T P.800 methodology. Both a DCR (Degradation Category Rating) and an ACR (absolute category rating) test were run at an independent test lab. Both the DCR and ACR tests were run using 32 listeners, 36 conditions, and 192 votes per condition.

The results from the DCR test are shown in Figure 4, with error bars indicating 95% confidence intervals. Note that BBE1 was not included in the test, as its subjective performance is very poor. It can be seen that the scores are consistent with the POLQA scores shown in Figure 3. The rank-order of the BBE algorithms is maintained, and BBE4 is again equivalent to AMR-WB at 8.85 kbps.

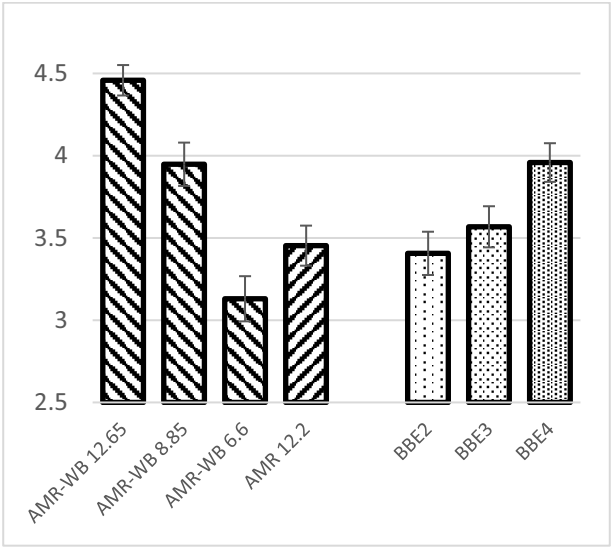

Figure 4: P.800 DCR MOS-LQS, at 3GPP mask level

The test results for the ACR are shown in Figure 5. It can be seen that the results are consistent with both the POLQA results, and the DCR results. Again, BBE4 matches AMR-WB 8.85's level of quality. The scores are shown in Table 1.

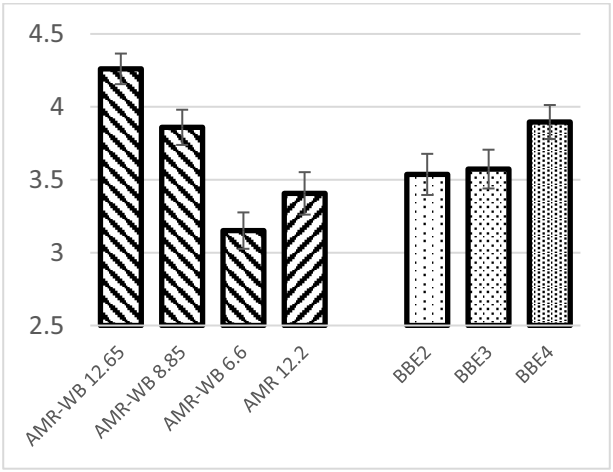

Figure 5: P.800 ACR MOS-LQS, at 3GPP mask level

Condition	DCR	ACR
AMR-WB 12.65	4.46	4.26
AMR-WB 8.85	3.95	3.86
AMR-WB 6.6	3.13	3.15
AMR 12.2	3.45	3.41
BBE2	3.41	3.54
BBE3	3.57	3.57
BBE4	3.96	3.90

Table 1: ACR vs DCR scores

4.4. Effect of high-band attenuation on subjective performance

In previous sections, it was suggested that BBE algorithms should be compared at a given bandwidth, and we suggest using the 3GPP WB Rx mask as the evaluation point, for maximum discrimination. However, it is not clear that this is the bandwidth that should be used in real-world deployments.

To establish this, we took our best performing algorithm, BBE4, tuned to meet the 3GPP WB Rx mask level, and applied several attenuations to the high-band, from 5 to 15 dB. This attenuation is denoted as N, as per Figure 2. Figure 6 shows the P.800 ACR and DCR scores for these conditions.

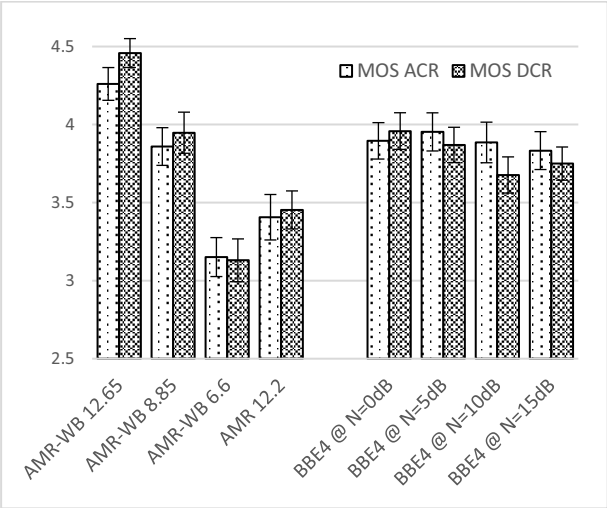

Figure 6: BBE4 ACR and DCR MOS vs bandwidth. N is the attenuation from the 3GPP WB Rx mask.

Several observations can be made. Firstly, there indeed appears to be an optimal operating point. For DCR, 0dB attenuation seems best. For ACR, 5 dB attenuation seems optimal. Note that these results must be taken with some degree of caution, as the differences observed are small, and not all statistically significant with 95% confidence.

This difference between ACR and DCR is expected: the DCR methodology presents the original wideband signal as a reference, therefore the results tend to weight bandwidth more, compared to an ACR test where the samples are presented without a reference. This can be tied to the observations from Section 3.3: the optimal operating point of BBE will probably be at a higher bandwidth if the network has both NB and WB, compared to a NB-only network.

Secondly, Figure 3 suggests that an optimal operating point for BBE4 would be around 5dB below 3GPP level, as POLQA score starts to drop above this point. This result matches the result of the ACR test, which is reasonable as POLQA is designed to predict ACR scores. Again, the objective methodology matches well with the subjective results.

4.5. Summary

Overall, results show that the proposed objective evaluation methodology, combining a POLQA score with a bandwidth requirement, works well. The results correlate well with both ACR and DCR testing, and in our testing, clearly identify which BBE algorithm performs best. In addition, it gives a good indication of the optimal level of bandwidth of a given algorithm.

It can also be noted that the best BBE algorithm we tested achieves quality equivalent to AMR-WB 8.85 when operating on AMR 12.2 transcoded inputs and meeting the 3GPP WB mask. This is consistent across testing methodologies, objective and subjective.

It can be argued that we have only tested a small number of BBE algorithms, and there is no guarantee that results will extend to all BBE algorithms. This is of course impossible to disprove, and is unavoidable considering the current limited number of BBE solutions commercially available in devices. However, even though the 4 BBE algorithms presented here use very different signal processing techniques, the conclusions have been consistent for all of them, giving confidence that they will extend to other BBE algorithms.

Previously, several papers have attempted to tackle the issue of objective vs subjective quality evaluation [11][12], but concluded that while there is reasonable correlation between objective and subjective scores, it is not reliable as a means to compare different BBE technologies. We believe that this may have been caused by not taking the bandwidth aspects into account. When considering the bandwidth, a reasonably reliable estimation of quality can be obtained.

5. Conclusions

In this paper, we present a methodology for objective evaluation of BBE algorithms, combining an objective metric with a bandwidth criterion. Results show that this methodology provides results consistent with that of subjective testing, both in terms of comparing different BBE algorithms, and comparing them to quality references such as AMR-WB 8.85. Additionally, the best algorithm tested here consistently matched the quality level of AMR-WB 8.85, and outperformed the AMR 12.2 narrowband input, according to all metrics tested.

We propose that this methodology be used by researchers for consistent BBE algorithm evaluation and comparisons, as well as by operators and terminal manufacturers for device testing. Finally, test results show that the use of good BBE provides a significant benefit over narrowband, and bridges the gap in quality between narrowband and wideband networks.

6. References

[1] 3GPP TS 26.190, “Adaptive multi-rate wideband (AMR-WB) speech codec; Transcoding functions,” 3rd Generation Partnership Project, Sept. 2012, version 11.0.0.

[2] 3GPP TS 26.441, “Codec for Enhanced Voice Services (EVS); General overview,” 3rd Generation Partnership Project, Dec. 2015, version 13.0.0.

[3] H. Carl and U. Heute, “Bandwidth enhancement of narrow-band speech signals,” in *Proc. EUSIPCO*, vol. 2, Edinburgh, UK, Sept. 1994, pp. 1178–1181

[4] H. Pulakka and P. Alku, “Bandwidth extension of telephone speech using a neural network and a filter bank implementation for highband mel spectrum,” *IEEE Trans. Audio, Speech, Language Process.*, vol. 19, no. 7, pp. 2170–2183, Sept. 2011

[5] 3GPP TS 26.131, “Terminal acoustic characteristics for telephony; Requirements,” 3rd Generation Partnership Project, Dec. 2015, version 13.2.0.

[6] ITU-T P.501, “Test signals for use in telephonometry,” Int. Telecommunication. Union, Jan. 2012

[7] ITU-T P.800, “Methods for subjective determination of transmission quality,” Int. Telecommunication Union, Aug. 1996

[8] ITU-T Rec. P.863, “Perceptual Objective Listening Quality Assessment,” Int. Telecomm. Union, Geneva, 2011

[9] ITU-T Rec. P.863.1, “Application guide for Recommendation ITU-T P.863,” Int. Telecomm. Union, Geneva, 2014

[10] 3GPP TS 26.090, “Adaptive multi-rate (AMR) speech codec; Transcoding functions,” 3rd Generation Partnership Project, Sept. 2012, version 11.0.0.

[11] Sebastian Möller et al, "Speech quality prediction for artificial bandwidth extension algorithms," in *INTERSPEECH 2013- 14th Annual Conference of the International Speech Communication Association, September 8–12, San Francisco, California, USA, Proceedings*, 2013 pp. 3439-3443.

[12] Pulakka, Hannu, Ville Myllylä, Anssi Rämö, and Paavo Alku. "Speech Quality Evaluation of Artificial Bandwidth Extension: Comparing Subjective Judgments and Instrumental Predictions," in *Interspeech 2015- 16th Annual Conference of the International Speech Communication Association*. 2015