SPEECH BANDWIDTH EXTENSION USING GENERATIVE ADVERSARIAL NETWORKS

Sen Li, Stéphane Villette, Pravin Ramadas, Daniel J. Sinder
Qualcomm Technologies, Inc., 5775 Morehouse Drive, San Diego, CA 92121-1714, United States
{senl, svillett, pramadas, dsinder}@qti.qualcomm.com

ABSTRACT

Speech blind bandwidth extension technologies have been available for some time, but until now have not seen widespread deployment, partly because the added bandwidth has been accompanied by added artifacts. In this paper, we present three generations of blind bandwidth extension technologies, from Vector Quantization mapping through Gaussian Mixture Models, to our latest architecture based on deep neural networks using Generative Adversarial Networks. This latest approach shows a sharp jump in quality, and demonstrates that machine-learning based blind bandwidth extension algorithms can achieve quality equal to wideband codecs, both objectively and subjectively. We believe that blind bandwidth extension can now achieve sufficiently high quality to warrant deployment in the existing telecommunication networks.

Index Terms— blind bandwidth extension, artificial bandwidth extension, generative adversarial network, 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 an 8 kHz sampling rate being used and in a practical frequency range of 300 – 3400 Hz. 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" — the use of wideband (WB) or super-wideband (SWB) coders, respectively, which use sampling rates of 16 kHz or 32 kHz and correspond to frequency ranges of 50 – 7000 Hz or 50 – 14000 Hz, respectively [1] [2].

However, WB and SWB deployments are not ubiquitous, as there can be substantial costs to develop, test, and deploy the supporting services. Further, end-to-end WB/SWB calls require upgraded devices at both ends. It will likely take years before full coverage and complete handset penetration is achieved, and upgrading landline networks to WB/SWB is likely to take even longer. Until then, a significant proportion of calls will still use legacy narrowband.

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.

2. BACKGROUND

2.1. Related work

Various statistical approaches to BBE have been proposed, to predict the 4-8 kHz portion of speech, usually referred to as the high-band (HB), from the 0-4 kHz portion, known as the low-band (LB). Typically, some form of either spectral folding or statistical modelling is used to generate a signal having the general characteristics of wideband speech [3] [4]. While perfect prediction cannot be expected, reasonably high quality speech can be obtained.

Vector Quantization (VQ) codebook mapping can be used to create discrete mapping of speech parameters from LB to HB [5][6]. Gaussian Mixture Models (GMM) based methods are used to preserve a more accurate transformation between LB and HB by modeling the speech envelope parameter continuously [7]. Hidden Markov Models (HMM) extend GMMs by exploiting speech temporal information [8]. Neural network based approaches such as deep neural networks have been proposed for BBE, as they are known to better model highly non-linear problems [9].

2.2. Loss functions and GANs

The statistical models discussed above are all based on the most basic loss function in regression problems – Mean Squared Error (MSE), which measures the difference in HB speech envelope parameters between prediction and ground truth. The MSE loss function works well in the average sense, but struggles to handle the uncertainty inherent in recovering missing speech HB such as detailed spectral shape and voiced/unvoiced energy dynamics. Minimizing MSE encourages finding parameter-wise averages of plausible solutions which are typically overly-smooth and thus have poor perceptual quality.

GANs have been introduced in [10], and have been successfully used in image processing field, such as image-to-image translation [11], image super resolution [12], and text-to-image synthesis [13]. The GAN training procedure encourages the reconstructions to move towards regions of the search space with high probability of containing realistic HB speech parameter distribution and thus close to the natural speech HB manifold [12]. In this paper we investigate how GANs may help for BBE.

3. BBE FRAMEWORK

In general, BBE frameworks are based on the classic source filter speech production model. Using such a model, the wideband extension of the narrowband speech signal can be divided into two sub-tasks:

- Estimation of the high-band spectral envelope
- Extension of narrowband excitation signal

To synthesize the HB speech signal, we leveraged the HB model from the EVRC-WB [14]. Figure 1 shows the overall diagram of our BBE framework.

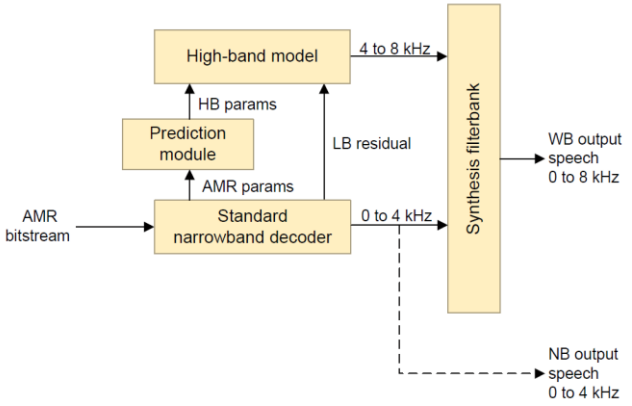

Figure 1: BBE framework

3.1. High-band Excitation

The HB excitation is derived from the NB excitation through a non-linear function, which generates a high-band excitation preserving the harmonic structure of the signal [14].

3.2. High-Band Spectral Envelope

In our speech HB extension model, for each 20 ms speech frame, 6th-order Line Spectral Frequencies (LSF) are used to spectrally shape the HB, together with a gain factor corresponding to the energy ratio between LB and HB [14].

3.3. Framework validation

This BBE framework has been tested to verify that it provides quality no worse than AMR-WB 12.65 kbps, both objectively and subjectively, when HB parameters are extracted from the original WB speech. Since BBE usually does not reach AMR-WB 12.65 kbps quality, the framework is not a performance bottleneck. This framework is also used in EVRC-WB, and the Qualcomm proprietary eAMR WB codec [16].

4. HB PARAMETER PREDICTION

4.1. Speech parameters

Input	Output
10th order low-band LSFs + Delta LSFs	6th order high-band LSFs
0-4 kHz speech energy	4-8 kHz speech energy

Table 1: Predictor input and output parameters

The parameters used in our HB prediction experiments are listed in Table 1. Backward deltas of the LB LSFs are used to improve the prediction without requiring extra delay.

4.2. Statistical Modeling with minimizing MSE

4.2.1. VQ Codebook Mapping

The most basic approach for BBE is codebook mapping. LB and HB speech envelope parameters are extracted from wideband speech and are further used to train a VQ codebook using a clustering method such as k-means. During the estimation phase, the received narrowband parameters are compared to the LB envelope parameter entries in the codebook, and the entry closest to the received narrowband envelope parameters is then chosen. The HB envelope parameters corresponding to the selected entry are used as the HB spectral envelope parameters [5]. In practice, the N-closest codebook entries are interpolated, weighted by the distance between their LB envelope parameters and the received narrowband envelope parameters [6].

4.2.2. Gaussian Mixture Models (GMM)

Compared to codebook mapping, GMM can model the speech envelope data continuously, which allows for soft clustering. Training is performed using Expectation Maximization (EM) and Maximum Likelihood Estimation (MLE) [7]. This probabilistic framework also has the flexibility to incorporate speech temporal information by introducing the state transition probability matrix during training, which converts the model to a GMM/HMM hybrid model. The main benefit from adding an HMM component is that it can implicitly exploit information from preceding speech frames to improve the estimation accuracy [8]. Morphing techniques from LB parameters to HB parameters using mixture means and covariance matrices are discussed in detail in [7].

4.3. Statistical Modeling with GANs

4.3.1. Generative Adversarial Networks framework

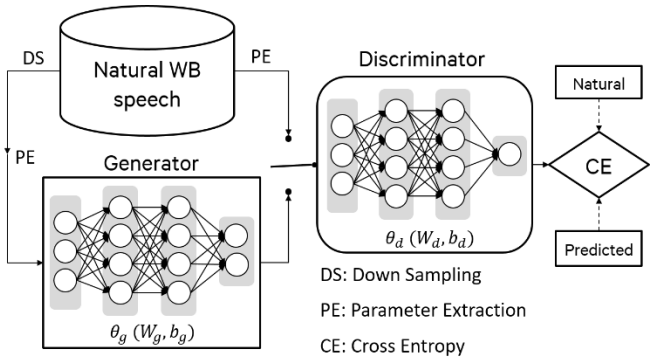

Figure 2: BBE-GAN Framework

A GAN [10] is comprised of a generator (G) and a discriminator (D), as shown in Figure 2. Here, for our BBE-GAN system, G is a deep neural network which predicts the

HB parameters from the LB parameters. D is another deep neural network acting as a binary classifier, which tries to differentiate between predicted HB parameters and natural HB speech parameters.

During adversarial training, G tries to fool D by adapting its weights and biases so that D believes its output is natural. D and G are iteratively trained, each trying to defeat the other. This approach leads G to generate an output that follows the same distribution as the natural data, and therefore can lead to more natural sounding speech.

4.3.2. Pre-training with MSE loss

Deep Neural Networks have been previously applied to the BBE problem, using an MSE loss, e.g. in [9]. We use such a model as a starting point. Here, a four-layer DNN generator of HB LSFs and energy is pre-trained using the standard MSE loss. This pre-training stage is crucial so that the GAN training process starts with a good initial generator, which helps avoid instability issues.

4.3.3. Perceptual loss function

The definition of our perceptual loss function l is critical for the performance of the generator network. Inspired by the perceptual loss function design in SRGAN [12], we combined the HB speech envelope parameter domain MSE l_{params} with the adversarial loss l_{adv} together and formulated the perceptual loss as their weighted sum, as per (1).

$$l = l_{params} + 10^{-2} * l_{adv} \qquad (1)$$

5. EXPERIMENTS

5.1. Setup

We conducted our speech bandwidth extension experiments using the NTT 1994 multi-language corpus [17] as our training and validation data with a 10-fold cross validation scheme. The data is sampled at 16 kHz sampling rate and digitized into 16-bit resolution, and an ITU-T P.341-compliant filter is applied to simulate a typical Tx handset response. We use ITU-T P.501 British English [18] as the evaluation dataset.

For BBE-VQ, we used separate 256-element VQ codebooks for HB LSFs and Gain. A weighted combination of the three closest candidates is used for the prediction.

For BBE-GMM, we used a GMM + HMM hybrid model with 64 states with 4 mixtures per state, and full covariance matrices. The forward path of the Viterbi decoding algorithm is used, i.e. no look-ahead delay is needed.

For BBE-GAN, both generator and discriminator are a four-layer feed-forward DNN (1 input layer, 1 output layer, 2 hidden layers) with 1024 neurons per hidden layer. ADAM optimizer is used during training.

Figure 3 and 4 show the spectral envelope of a typical voiced segment and an unvoiced segment during the adversarial training process at iteration 0, 100 and 200. We can clearly see that moving away from MSE as the loss function, BBE-GAN output is moving towards the spectrum of the reference WB speech. The GAN training process is seen to improve the energy of unvoiced segments while cleaning up unwanted HB noise during voiced segments. This leads to a noticeable increase in speech quality, with less audible artifacts and higher naturalness.

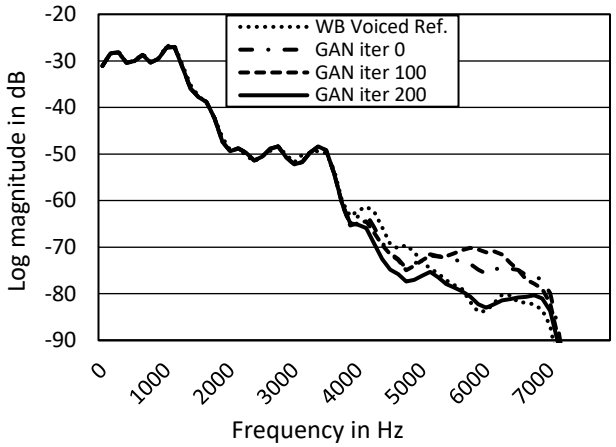

Figure 3: Voiced speech output vs GAN iterations

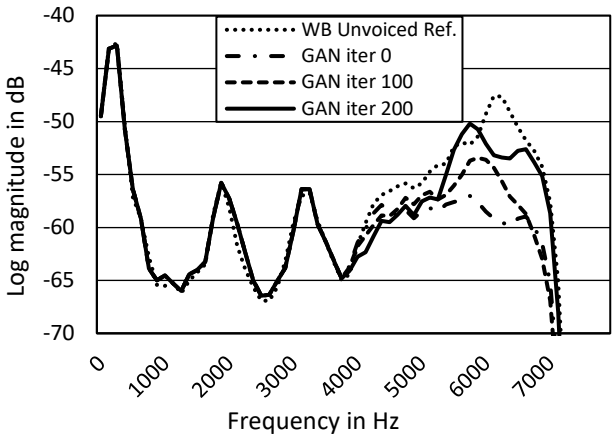

Figure 4: Unvoiced speech output vs GAN iterations

5.2. Objective performance

For objective evaluation, we followed the methodology described in [19], and defined in the ITU-T P Suppl. 27 [20]. For the bandwidth requirement, we measure the Rx frequency response respective to the 3GPP Rx mask [21], using ITU-T P.501 British English speech material as the input. For speech quality, we measure the POLQA [22] score of BBE output with P.501 British English coded by AMR 12.2 kbps.

We plotted the POLQA scores for the BBE algorithms discussed above. The scores for AMR-NB at 12.2 kbps and AMR-WB at 8.85 kbps and 12.65 kbps are shown as the references. The results are shown in Figure 5, where 0dB indicates the response follows the lower limit of the mask. There is clear improvement from BBE-VQ to BBE-GMM to BBE-GAN, showing the increasing modeling power of the

statistical models used. Between GAN at iteration 0 and GAN at iteration 200 (fully trained), the maximum POLQA value is similar, however BBE-GAN at 200 iterations does maintain its POLQA score better at higher amounts of bandwidth. This is a good indication of the prediction quality, and is made possible by the reduction in the number of prediction artifacts from the fully trained GAN.

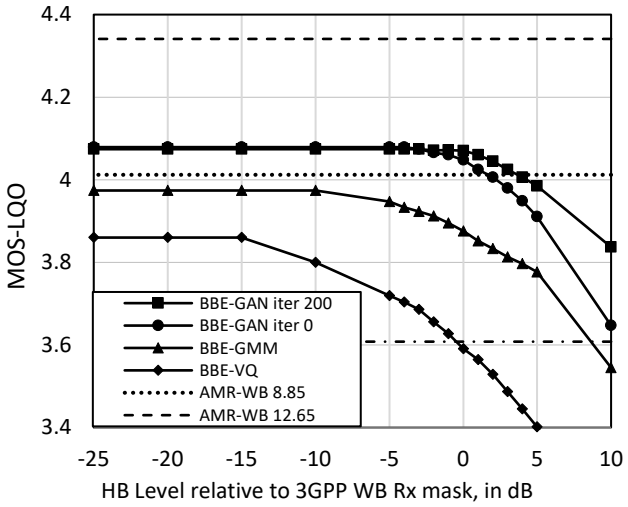

Figure 5: POLQA MOS-LQO vs Bandwidth

5.3. Subjective performance

The subjective performance of the various BBE algorithms presented here was evaluated using the ITU-T P.800 methodology. A Degradation Category Rating (DCR) [23] test was run at an independent test lab. The test was run using 32 listeners, 42 conditions and 192 votes per condition. The results from the DCR test are shown in Figure 6, with error bars indicating 95% confidence intervals. The scores are consistent with the objective results shown in Figure 5.

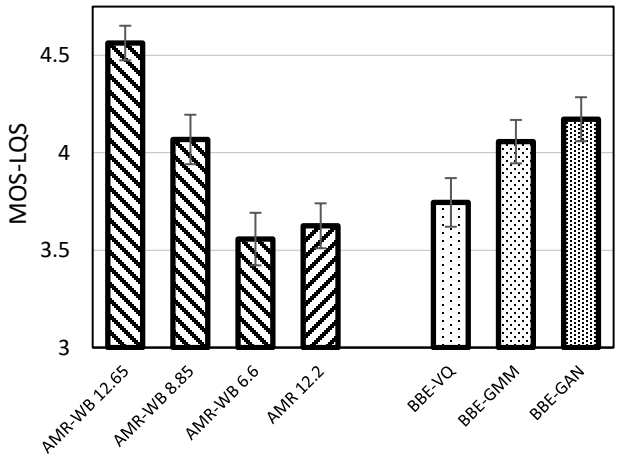

Figure 6: P.800 DCR MOS-LQS at 3GPP mask level

The rank-order of the BBE algorithms is maintained and BBE-GAN is statistically equivalent to AMR-WB at 8.85 kbps. More testing results can be found for BBE-VQ and BBE-GMM in [19] (where they respectively correspond to algorithms BBE3 and BBE4).

5.4. HB attenuation vs subjective quality

We applied several filters to BBE-GAN to adjust the HB level from +5dB to -10dB relative to the 3GPP WB Rx mask. Figure 7 shows the P.800 DCR scores for these conditions. Note that, as in Figure 5, the level is relative to the lower mask limit, so that -5dB indicates a response below the lower limit of the mask, whereas +5dB indicates a response between the upper and lower limits of the mask.

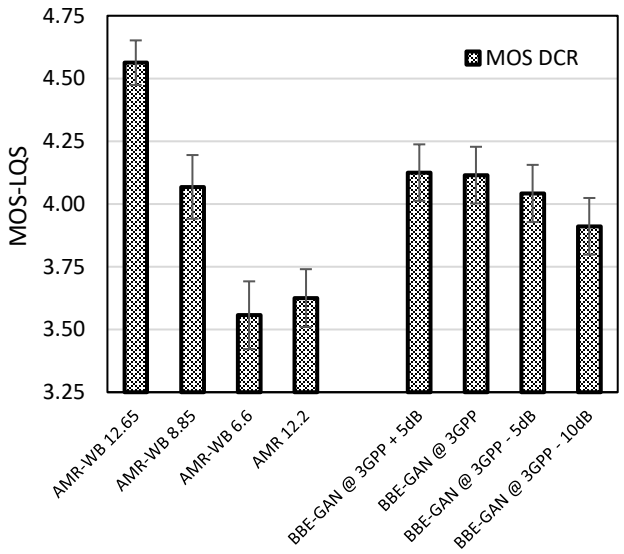

Figure 7: DCR MOS vs bandwidth

We observed that BBE-GAN maintains performance even at higher bandwidth levels, as predicted by the objective metric results shown in Figure 5. This also suggests that BBE-GAN is fully comparable with WB codec both in terms of bandwidth and quality, and confirms again that the objective evaluation aligns well with subjective results [19][20].

6. CONCLUSION

In this paper, we presented three generations of blind bandwidth extension technologies, from VQ to GMM to GAN. We find that machine learning such as GAN allow a significant step up in quality compared to classic statistical modeling techniques. GAN-based prediction allows the quality of BBE to be similar to WB codecs, achieving performance equivalent to AMR-WB 8.85 kbps quality both objectively and subjective. While BBE technology has been studied for many years, it has not been widely deployed as it could not offer quality similar to that of wideband codecs. We have shown that the use of machine learning techniques such as GAN allows BBE to reach that level of quality, which may potentially accelerate widespread adoption of BBE in telecommunication networks.

7. 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] Y. Qian and P. Kabal, "Wideband speech recovery from narrowband speech using Classified codebook mapping", Proceedings of the 9th Australian International Conference on Speech Science & Technology Melbourne, Dec. 2002.

[6] J. Epps and W. H. Holmes, "A new technique for wideband enhancement of coded narrowband speech," in Proc. IEEE Speech Coding Workshop, 1999, pp. 174–176.

[7] K.-Y. Park and H. S. Kim, "Narrowband to wideband conversion of speech using GMM based transformation," in Proc. ICASSP 2000, pp.1843–1846.

[8] P. Jax and P. Vary, "Artificial bandwidth extension of speech signals using MMSE estimation based on a Hidden Markov model," in Proc. ICASSP 2003, pp. 680-683.

[9] Y. Wang, S. Zhao, W. Liu, M. Li, J. Kuang, "Speech bandwidth expansion based on Deep Neural Networks," in Proc. INTERSPEECH 2015, pp. 2593-2597.

[10] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. "Generative adversarial nets," in Advances in Neural Information Processing Systems (NIPS), pages 2672–2680, 2014.

[11] P. Isola, J. Zhu, T. Zhou, A. A. Efros, "Image-to-Image Translation with Conditional Adversarial Networks," arXiv:1611.07004.

[12] C. Ledig, *et al.* "Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network," arXiv:1609.04802.

[13] H. Zhang, *et al.* "StackGAN: Text to Photo-realistic Image Synthesis with Stacked Generative Adversarial Networks," arXiv:1616.03242.

[14] 3GPP2 C.S0014-C v1.0 "Enhanced Variable Rate Codec, Speech Service Option 3, 68 and 70 for Wideband Spread Spectrum Digital Systems".

[15] 3GPP TS 26.090, "Adaptive multi-rate (AMR) speech codec; Transcoding functions," 3rd Generation Partnership Project, Sept. 2012, version 11.0.0.

[16] S. Villette, S. Li, P. Ramadas, D. Sinder, "eAMR: Wideband speech over legacy narrowband networks," in Proc. ICASSP 2017, pp. 5110-5114.

[17] N. A. T. Corporation, "Multi-lingual speech database for telephonometry," http://www.nttat. com/products e/speech, 1994.

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

[19] S.Villette, S. Li, P. Ramadas, D. Sinder, "An Objective Evaluation Methodology for Blind Bandwidth Extension," in Proc. INTERSPEECH 2016, pp 2548-2552.

[20] ITU-T P Suppl. 27, "Application of ITU-T P.863 and ITU-T P.863.1 for speech processed by blind bandwidth extension approaches," Int. Telecomm. Union, Geneva, 2017.

[21] 3GPP TS 26.131, "Terminal acoustic characteristics for telephony; Requirements," 3rd Generation Partnership Project, Dec. 2015, version 13.2.0.

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

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