Bandwidth extension method based on generative adversarial nets for audio compression

Qingbo Huang, Xihong Wu, and Tianshu Qu

Key Laboratory on Machine Perception (Ministry of Education), Speech and Hearing Research Center, Peking University, China

Correspondence should be addressed to Tianshu Qu (qutianshu@pku.edu.cn)

ABSTRACT

The compression ratio of core-encoder can be improved significantly by reducing the bandwidth of the audio signal, resulting in poor listening perception. This paper proposes a bandwidth extension method based on generative adversarial nets (GAN) for extending the bandwidth of an audio signal, to create a more natural sound. The method uses GAN as a generative model to fit the distribution of the MDCT coefficients of the audio signals in the high-frequency components. Through minimax two-player gaming, more natural high-frequency information can be estimated. On this basis, a codec system is built up. To evaluate the proposed bandwidth extension system, MUSHRA experiments were carried on and the results show that there is comparable performance with HE-AAC.

1 Introduction

For the convenience of storage and transmission, the audio signals are usually compressed through a lossy single-channel codec. At the low bitrate, the high-frequency components are truncated at the encoder, leading to some uncomfortable feelings like “muffled”, “dull” etc. To solve the above problems, the bandwidth extension technology comes into being. The bandwidth extension technology refers to the regeneration of the high frequency part from the low frequency part in the decoder with or without prior information in order to improve the sound quality, making the decoder’s output “warm” and “sparkle”.

Bandwidth extension methods are well studied. The core problems are the modelling of the relationship between the high frequency part and the low frequency part. According to how to representation the relationship, the bandwidth extension methods falling into two categories, rules based and generative models based.

The salient examples in the first category are as follows: in 1979, Makhoul and Berouti proposed a method to expand the bandwidth of voice signals by spectral folding and shifting [1]. This method is generally considered as the first method of audio bandwidth extension technique. In 1997, Coding Technology proposed the Spectral Band Replication (SBR) which successfully applied psychoacoustic model [2-3] as an evaluation criterion [4]. By virtue of its excellent performance, the SBR becomes an important module in HE-AAC. In 2002, Chung et al. exploits the mutual information between bands, including gain, spectral envelope, and harmonic [5]. Also in 2002, Lyubimov and Lukin extended the bandwidth according to the harmonic relationship [6]. Seo proposed a bandwidth extension method in Bark domain [7] in the same year. In 2009, Nagel and Disch proposed harmonic band extension method [8].
The rules in the band extension method are often hard to summarize under complicated circumstances, so generative models are introduced to deal with the problem. The generative models are divided into the explicit density models and the implicit density models. The explicit density models need the probability density function which is often hard to get. According whether the probability density function is tractable, the explicit density models can be divided into the tractable density models and the approximate density models. For the tractable density models, in 1994, Cheng et al. proposed the method using the Statistical Recovery Function (SRF) to model the relationship of the AR model’s coefficients between the low frequency part and the high frequency part [9], which is a salient example. If we could get the exactly density function, the tractable density model is the best generative model. However this assumption does not hold in most cases.

So through training or learning to get an approximate density function may be a solution. All the method based on that idea is the approximate density models, such as Hidden Markov Model (HMM), Gaussian Mixture Model (GMM) and so on. The examples are as follow: in 2000, Jax and Vary used the HMM to accomplish the voice bandwidth extension [10]. In the same year, Park et al. used the GMM to extend the voice bandwidth [11]. In 2009, Lyubimov and Lukin used the Cluster Weighted Model to model the Mel-Frequency Cepstral Coefficients (MFCC) to extend the audio bandwidth [12]. In 2015, Chung used the GMM to extend the audio bandwidth [13]. Although the approximate density model is widely used and achieves success, there are always some unexplainable assumptions, such as how many Gaussian density functions should be used in the GMM model.

Recently, with the rapid development of the neural network, great success had been made in the audio bandwidth extension technology area. In 2010, Pham et al. used Feed Forward Neural Network (FFNN) to extend the voice bandwidth [14]. In 2011, Liu et al. used Radial Basic Function (RBF) neural network to extend the voice bandwidth [15]. In 2012, Pulakka and Alku used neural networks to extend the spectrum based on narrow-band speech features [16]. In 2015, Liu proposed a bandwidth extension of audio signals by using a similarity correlation degree-based neural network [17].

In the implicit density models, prior assumptions become weak at the expense of the intractable probabilistic computations are unacceptable. The Generative Adversarial Networks (GAN) [18] solves this problem. As a generative model, GAN has the advantages that use the discriminative model to evaluate the generative model, reducing the computation complexity by using an indirect method.

For this reason, GAN is widely used in image compression. Larsen et al. combine GAN with an autoencoder. The autoencoder is used as the generation network G to compress the coding of the image, and to evaluate the result [19] after the codec with the discriminant network D. Another application for GAN is image super-resolution, using a series of convolutional networks in Laplacian pyramid framework to generate images from rough to fine [20].

Referring to the related work, this paper proposes an bandwidth extension method based on GAN for audio compression, looking forward to get a more natural result through introducing GAN as the generative model. The paper is organized as follows: the bandwidth extension method based on generative adversarial nets is presented in Section 2 and the GAN based bandwidth extension codec is further described in Section 3. The subjective quality test results are presented in Section 4 and in section 5, we draw the conclusions.

2 The bandwidth extension method based on generative adversarial networks

In 2014, Goodfellow proposed the GAN. The GAN pits two networks against each other: a generative model $G$ that captures the data distribution and a discriminative model $D$ that distinguishes between samples generated by $G$ and that from the training data. In our approach, both $G$ and $D$ are feed
forward neural networks. The generative network $G$ takes a noise vector $z$ drawn from a distribution $p_{\text{Noise}}(z)$ as input and outputs a high frequency spectrum $\tilde{h}$. The discriminative network $D$ takes a frequency spectrum as input, which is randomly to be either $\tilde{h}$ generated by $G$, or a real frequency spectrum $h$ drawn from the training data $p_{\text{Real}}(h)$. $D$ outputs a scalar probability, which is trained to be “1” if the input was real and “0” if generated from $G$. A minimax objective is used to train both networks together:

$$\min \max V(D, G) = E_{h \sim p_{\text{Real}}(h)}[\log D(h)] + E_{z \sim p_{\text{Noise}}(z)}[\log(1 - D(G(z)))]$$

(1)

where $E$ is the expectation function, $V$ is the loss function. This encourages $G$ to fit $p_{\text{Real}}(h)$ so as to fool $D$ with its generated samples $\tilde{h}$. Both $G$ and $D$ are trained by back-propagating the loss in equation (1) through their respective models to update the parameters.

The Conditional Generative Adversarial Net [21] (CGAN) is an extension of the GAN where both networks $G$ and $D$ receive an additional vector of information $l$ as input. This might contain, say, information about the class of the training example $h$. The loss function thus becomes,

$$\min \max V(D, G) = E_{h \sim p_{\text{Real}}(h)}[\log D(h, l)] + E_{z \sim p_{\text{Noise}}(z)}[\log(1 - D(G(z, l), l))]$$

(2)

where $p(l)$ is, for example, the prior distribution over classes. This model allows the output of the generative model to be controlled by the conditioning variable $l$. Mirza and Osindero explored this model [21] with experiments on the Modified National Institute of Standards and Technology (MNIST) database and faces, using $l$ as a class indicator. In our approach, $l$ will be the high frequency spectrum from the training dataset. According to the characteristics of the audio bandwidth extension task, the spectrum envelope of the high-frequency part are introduced to GAN as a priori information to ensure that the output of the $G$ is consistent with the real data in spectrum envelope.

$$\min \max V(D, G) = E_{h \sim p_{\text{Real}}(h)}[\log D(h, l)] + E_{z \sim p_{\text{Noise}}(z)}[\log(1 - D(R(G(z, l), l)))]$$

(3)

where $R$ is the revised function to make the networks’ output keeping the same envelope as the original.

Figure 1. GAN training framework.

$$E_r(n) = \min_{k \leq \text{cutf\_slen}} \max_{1 \leq n \leq \text{cutf\_slen}} (\text{MDCTcoef}[(k)])^2, 1 \leq n \leq \text{cutf\_slen}$$

(3)

$$E_r(n) = \min_{k \leq \text{cutf\_slen}} \max_{1 \leq n \leq \text{cutf\_slen}} (\text{MDCTcoef}[(k)])^2, 1 \leq n \leq \text{cutf\_slen}$$

(4)

$$E_r(n) = \frac{E_r(n)}{E_r(\text{cutf\_slen})}, 1 \leq n \leq \text{cutf\_slen}$$

(5)

Corrected Data(k) = Generated Data(k) × \left(\frac{k}{\text{cutf\_slen}}\right)$$

(6)

The GAN training flow is shown in Figure 2: $G$’s input is low-band data, and the generated high-frequency data is modified by the module to obtain corrected high-frequency data according to prior information according to equation (6). $E_r$ is calculated according to equation (3-5), where slen is the subband length and the cutf is the cutting-off frequency. The corrected high frequency data is then combined with the low frequency data to generate the final fake data. The corresponding original high-frequency part and low-frequency part synthesis the true data.

To avoid pre-echo, the transient parts of the audio signals need to be detected. For transient parts, the short frame is applied to get higher time resolution. The long frame is chosen for steady-state parts. Moreover, the distributions are different in transient signals and in steady-state signals. Therefore, two GANs are trained respectively.
3 GAN based bandwidth extension system

Based on the proposed bandwidth extension method in section 2, a complete codec system is set up with the MPEG-4 AAC Low Complex as the core encoder.

The encoding process is shown in Figure 2. The audio signal in time-domain is put into transient detection module to decide the frame type, and mark it as the type symbol. The steady-state signal uses long frames, while the transient signals use short frames. After MDCT transformation, the MDCT spectrum is divided into the high frequency part and the low frequency part according to the cutting-off frequency in network training. The energy ratio of high frequency and low frequency energy is calculated and quantified. The calculation method is the same as that in network training. The quantized envelope energy ratio, the frame type mark and the core encoder result are synthesized into the bit stream.

![Figure 2. Proposed encoding framework.](image)

The encoding process is shown in Figure 3. Firstly, extract the core codec’s coding result, the envelope energy ratio and the frame type symbol from the bit stream. The core codec’s bit steam passes the core decoder to obtain the low-frequency signal in the time domain. The low-frequency signal performs the corresponding long or short MDCT transform according to the frame type symbol. The low-frequency spectrum passes through the generation network in the GAN. The output of the network is tuned by the correction module according to the high-low frequency envelope energy ratio to obtain the generated high frequency spectrum. The high frequency spectrum is performs the corresponding long or short IMDCT transform according to the frame type symbol to get the high-frequency signal in time-domain. At last, the low-frequency signal and the high-frequency signal in time-domain are synthesized to get the final wide band audio signal.

4 Subjective evaluation experiments

For steady-state signals, we perform 2048 point MDCT and set the 16 spectral lines to a subband to calculate the energy. The structure of the $G$ net in GAN is shown in Figure 4. All the activation functions are the hyperbolic tangent functions.

The structure of the $D$ net in GAN is shown in Figure 5. In the hidden layers and the input layer the activation functions are the hyperbolic tangent functions and in the output layer the activation function is the sigmoid functions.

![Figure 4. $G$ net structure for steady signal.](image)
output layer is 16 in $G$ net. The neuron units’ number in input layer in $D$ net is 16.

Training database is 60 min music and speech. To monitor the convergence of the network, the output of $G$ is saved as the intermediate results. The black line is the output MDCT spectrum of the generative net after specific epochs training and the red dotted line is the ground truth. Figure 6-8 is the output MDCT spectrum of the $G$ nets after 20, 100, 200 epochs training respectively. It is obvious that the black line is getting more and more similar to the red dotted line as training epoch grows.

Figure 5. $D$ net structure for steady signal.

Figure 6. The output MDCT spectrum of the $G$ nets after 20 epochs training.

Figure 7. The output MDCT spectrum of the $G$ nets after 100 epochs training.

Figure 8. The output MDCT spectrum of the $G$ nets after 200 epochs training.

To evaluate the proposed bandwidth extension system, MPEG-4 HE-AAC is chosen as the baseline due to the proposed codec share the same core codec as MPEG-4 HE-AAC. The bit rate of the core codec is set as 30kbps and the bit rate of the bandwidth extension module’s side information is set as 2kbps.

According to the subjective evaluation method for intermediate quality level of audio systems specified by the International Telecommunication Union, Multiple Stimuli with Hidden Reference and Anchor [22] (MUSHRA) experimental paradigm is recommended to evaluate the sound quality of the proposed codec. The audio signals from MPEG data set are selected for the evaluation experiment, and all these audio signals are sampled at 44100Hz and
other detail information are shown in the following table. 12 students (6 males and 6 females) with normal hearing aged 22-27 years old participated in the experiment as subjects. The subjective experiment were performed in a quiet listening room where the reverberation time above 100Hz is less than 300ms and the background noise is 20.9dB and choosing Sennheiser HD650 headphone as a playback device.

Table 1. MPEG verification test files of 1997

<table>
<thead>
<tr>
<th>Type</th>
<th>file name</th>
<th>Content types</th>
<th>duration (s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>speech</td>
<td>es01</td>
<td>Suzanne Vega</td>
<td>10.734</td>
</tr>
<tr>
<td></td>
<td>es02</td>
<td>Male German Speech</td>
<td>8.599</td>
</tr>
<tr>
<td></td>
<td>es03</td>
<td>Female English</td>
<td>7.604</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Speech</td>
<td></td>
</tr>
<tr>
<td>multi-instrument</td>
<td>sc01</td>
<td>Trumpet</td>
<td>10.968</td>
</tr>
<tr>
<td></td>
<td>sc02</td>
<td>Orchestra</td>
<td>12.732</td>
</tr>
<tr>
<td></td>
<td>sc03</td>
<td>Contemporary Pop</td>
<td>11.552</td>
</tr>
<tr>
<td>single instrument</td>
<td>si01</td>
<td>Harpsichord</td>
<td>7.995</td>
</tr>
<tr>
<td></td>
<td>si02</td>
<td>Castanets</td>
<td>7.725</td>
</tr>
<tr>
<td></td>
<td>si03</td>
<td>Pitch pipe</td>
<td>27.887</td>
</tr>
<tr>
<td>instrument solo</td>
<td>sm01</td>
<td>Bagpipe</td>
<td>11.148</td>
</tr>
<tr>
<td></td>
<td>sm02</td>
<td>Glockenspiel</td>
<td>10.095</td>
</tr>
<tr>
<td></td>
<td>sm03</td>
<td>Plucked Strings</td>
<td>13.985</td>
</tr>
</tbody>
</table>

The MUSHRA scores for average and 95% confidence intervals are presented in Figure 9-12.
Using SPSS to analyse the MUSHRA score, it is generally considered that there are significant differences between the two systems when \( p<0.05 \). Overall, the new system is almost indistinguishable from HE-AAC. The proposed system is better than HE-AAC for files sc01, sc02, sc03, si03 and sm01 but not significant. For the files es01, es02, si01, si02, sm02, sm03, HE-AAC is better than the proposed system, and for files es01, si02, sm02, sm03, the difference between the two systems is significantly.

Generally speaking, the proposed method is comparable with the SBR module in HE-AAC in most cases. For the files es01, si02, sm02, sm03, HE-AAC performs better than the proposed system significantly. As shown in table 1, these 4 files are all transient signals. Although transient signals is detected and processed specially, the result shows the proposed method performs inferiorly in transient signal. This may be cause by that for transient signals the relationship such as harmonic structure between the low frequency part and the high frequency part is not stable and difficult to model.

5 Conclusion

In this paper, we proposed a GAN based bandwidth extension method for audio compression. A new bandwidth extension method based on GAN was proposed. A complete audio coding system was built by using AAC-Low Complex as the single-channel core encoder and using the proposed bandwidth extension method to regenerate the high frequency part. The new system fixes a series of discomforts caused by the core codec at low bit rate where the cutting-off frequency is too low. In order to evaluate the audio quality decoded by the new system, a subjective evaluation experiment was carried out using the HE-AAC as the baseline system by using the MUSHRA experimental method. Through statistical analysis of the experimental results, the new system proposed in this paper has no significant difference with HE-AAC.

6 Acknowledgment

This work was supported by the National High Technology Research and Development Program of China (863 Program: 2015AA016306), and the National Natural Science Foundation of China No. 61175043, No.61421062, No. 31170985), and supported by High performance Computing Platform of Peking University.

References


