LPC-GAN for Speech Super-Resolution

Konstantin Schmidt, Bernd Edler
International Audio Laboratories Erlangen
Am Wolfsmantel 33 91058 Erlangen, Germany
konstantin.schmidt@audiolabs-erlangen.de

Ahmed Mustafa Mahmoud, Guillaume Fuchs
Fraunhofer IIS
Am Wolfsmantel 33 91058 Erlangen, Germany

***Abstract*—Up to today telephone speech lacks of perceptual quality and intelligibility due to bandwidth removal and quantisation artefacts in the encoding process. Super-resolution artificially regenerates this missing frequency content and thus improves the perceptual quality and intelligibility. This work proposes a novel approaches for super-resolution based on generative adversarial networks with convolutional architectures. Motivated by the source-filter model of the human speech production, the proposed system decomposes the speech signal into spectral envelope and excitation signal. The missing frequency-content of envelope and excitation are restored with dedicated networks. The network restoring the excitation signal is trained such that there is no mismatch between the excitation signal and the envelope. By this, we achieve better perceptual quality at lower computational complexity.**

***Index Terms*—speech enhancement, speech super-resolution, bandwidth extension, artificial bandwidth expansion, audio super-resolution**

I. Introduction

Speech communication is a technology used by most people every day, creating a vast amount of data that needs to be transmitted over Voice over IP, cellular or public switched telephone networks. While the amount of transferred data should be kept low, the quality of speech is desired to be high. In order to reach this goal, speech compression technologies have evolved over the past decades from compressing bandlimited speech with simple pulse code modulation to coding schemes following speech production and human perception models able to code fullband speech. Albeit the existence of such standardised speech codecs, their adoption in cellular or public switched telephone networks takes years if not decades. For this reason AMR-NB [1] remains the most frequently used codec for mobile speech communication which merely encodes frequencies from 200 Hz to 3400 Hz (usually named *narrowband*, NB). However, transmitting band-limited speech not only harms the acoustic quality but also the intelligibility [2], [3]. Super-resolution (SR) artificially regenerates missing frequency components without transmitting additional information from the encoder. A SR can be added to the decoder toolchain without any adaption of the transmission network and thus can serve as an intermediate solution to improve the perceptual audio quality and intelligibility until better codecs will be deployed in the network.

This work presents a SR based on deep convolutional networks using adversarial learning targeting speech coding scenarios. To summarise our contribution, we show how the well-known separation of speech signals into excitation signal and envelope can be implemented with deep convolutional architectures and we successfully apply it to SR. By this we can reduce the computational complexity while increasing the perceptual quality. We train all networks with a mixture of adversarial and Mel-loss. This allows for perceptually motivated loss while retaining the advantages of adversarial loss.

II. State-of-the-Art

Early SRs already utilised the separation of the speech signal into excitation and spectral envelope. These systems apply statistical models to extrapolate the spectral envelope while generating the excitation signal by spectral folding [4], spectral translation [5] or by nonlinearities [6]. Such statistical models are Hidden Markov Models [4] or later DNNs [5], [7]–[12].

Since artificial excitation generation introduces artifacts, it may be beneficial to extrapolate the time-domain signal by DNNs. Unfortunately the probability distribution of time-domain speech is very complex and hard to model, even with today's powerful networks. Models trained with L^p-loss or cross-entropy loss to match this complex distribution, will only produce a smoothed approximation thereof. When applied to SR, this means that the resulting speech signal will lack crispness and energy [13]. GANs [14] can be seen as a kind of extended loss function. Here, two networks, a generator and a discriminator compete against each other. The generator tries to generate realistic data while the discriminator distinguishes between the generated data and the data from the training database. After successful training, the discriminator is not needed any longer, its mere purpose lies in providing a better loss for the generator. The first SR extrapolating the time-domain signal was [9], the first using a GAN were [13], [15]–[18]. Other approaches than GANs to do SR are autoregressive networks [19].

III. Proposed Systems

The above mentioned SRs, that model the speech signal in time-domain, have the problem that the DNNs used for this task need to be quite large resulting in high computational complexity and memory requirements. In this work we show how speech signals separated into excitation and envelope can be modeled with GANs. Here the application is SR but not limited to it. The training objective used is a mixture of

adversarial and feature loss. Using this combination allows for simpler architecture and faster training of the whole system. For comparison we also present a second SR that models the speech signal without the separation into excitation signal and envelope. Both systems are trained with the same discriminator architecture, the same perceptual loss and the same optimisation algorithm. First, both generator networks are presented, the discriminator will be described at the end of this section. The input to all generator networks are raw time-domain speech samples coded with AMR-NB, the output are WB time-domain speech samples.

A. LPC-GAN

The proposed system is presented in Fig.1 where the input NB speech signal is separated into a set of LPCs representing the spectral envelope and an excitation signal. The excitation signal together with the input signal are fed to a first DNN for extrapolation to a WB excitation signal. This path operates on samples, shown here as solid lines. The LPCs are extrapolated to a WB envelope with a second DNN in the upper path. This path operates on frames of 15 ms, shown here as dashed lines. Since LPC coefficients are IIR filter coefficients and manipulations like extrapolation could result in an unstable filter, they are extrapolated in the LSF domain [20]. LSFs are a bijective transformation of LPCs with several advantages: First, they are less sensitive to noise disturbances and an ordered set of LSFs with a minimum distance between the coefficients will always guarantee a stable LPC filter. Second, the spectral envelope at a particular frequency depends mostly on one of the LSFs so an erroneous extrapolation of a single LSF coefficient mainly affects the spectral envelope at a limited frequency range. These properties make them suitable for being extrapolated to a set representing a WB envelope. The extrapolated LSF coefficients are transformed back to the LPC domain for shaping the extrapolated excitation signal, which forms the output signal. The extrapolated excitation signal, shaped by the LPC envelope, forms the output WB signal. When training the network extrapolating the excitation signal, any mismatch between the extrapolated envelope and excitation signal shall be avoided. For this reason the loss for training the DNN extrapolating the excitation signal is calculated on the shaped output speech and the gradient is propagated through the LPC filter. This can be achieved by implementing the LPC filtering as an additional DNN layer. Since the LPC filter is a all-pole IIR filter, this DNN layer should be a layer with recurrent units. Unfortunately, backpropagating gradients through a recurrent layer will cause the gradient to vanish [21] and result in poor training. As a solution to this problem, the IIR filter coefficients are transformed into FIR filter coefficients by calculating the truncated impulse response from the IIR filter. It is known from signal processing that any IIR filter can be approximated by an FIR filter by truncating the infinite impulse response [22]. Then, the LPC shaping can be implemented with a convolutional layer. Fig. 2 shows the effect of truncating it to 64 samples. While the IIR LPC envelope is smooth, the truncated FIR envelope has lots of ripples and does not follow well the IIR envelope in high frequencies. For this reason the LPC coefficients are multiplied with an exponential function before calculating the truncated impulse response $\hat{a}_i = a_i \cdot 0.8^i$. The resulting $\hat{a}_i$ coefficients have less pronounced poles and are suitable for calculating the FIR envelope as shown in Fig. 2. However, less pronounced poles result in less shaping and thus not being as efficient as all-pole IIR coefficients.

Initial experiments have shown that the FIR shaped signal contains artefacts, which could easily be identified by the discriminator. As a result, the adversarial loss was not balanced and the generator was training poor. This could be solved by calculating the adversarial loss on the real and generated *unshaped* excitation signal. The LPC shaping by an FIR filter is done only during training time. During evaluation time, no gradient needs to be backpropagated, so the LPC coefficients are applied as an IIR filter.

The DNN architecture used for extrapolating the excitation signal is a stack of convolutional neural layers (CNNs) similar to the previously published system in [23]. One of these layers is displayed in Fig. 3. Half of the output channels are fed into $tanh$-activations and the other half is fed into softmax activation. Both activations are multiplied over the channel dimension in order to form the output of each layer. There is also a residual connection from the input to the output in order to avoid vanishing gradients and maintain stable and effective training [24]. Our softmax-gated activations [23] are more efficient to compute, more robust against reconstruction

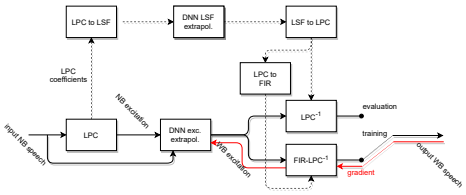

Fig. 1: Proposed system based on the decomposition of the speech signal into excitation signal and LPC envelope. All paths with solid lines operate in samples, all paths with dashed lines operate on frames of 15 ms. The path of the gradient during training is shown in red.

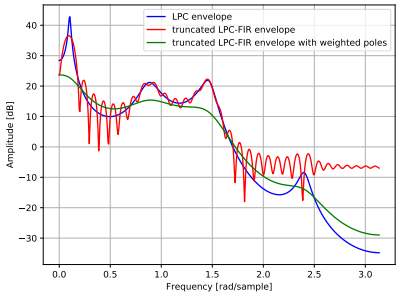

Fig. 2: Transfer functions of an IIR LPC filter of order 12 and FIR filters resulting from a truncated impulse response.

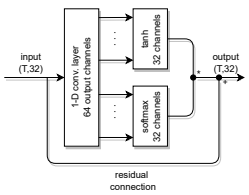

Fig. 3: Single layer of the generator network in the CNN-GAN and the LPC-GAN with softmax-gated activations.

artefacts and have shown faster convergence compared to the sigmoid-gated activations in WaveNet [25].

The weights of the convolutional kernels are normalised using weight normalisation [26] to enable stable training behaviour. We also apply batch normalisation to the output features from the CNN layers to speed up the training process. The additional frame-rate network extrapolating LPC coefficients in the LSF domain operates on the same gated CNNs described above, input and output are LSF coefficients, not time-domain samples. To add more time-context, a final GRU-layer has been added. To proof the benefits of LPC-GAN, we add a second system that extrapolates the speech signal directly without the proposed separation into excitation and envelope. It is based on the same convolutional architecture described in III-A only with more kernels. The reason for the increased model size is that initial experiments with smaller models were not performing acceptable for a listening test. This system is denoted CNN-GAN. A stable adversarial training is achieved by applying spectral normalisation to the convolution kernels of the discriminator network [27]. This kind of normalisation enforces the Lipschitz condition to the function learned by the discriminator, which was found important for an effective and stable adversarial training procedure. The discriminator operates in conditional setting [28], hence the input signal includes the real or fake WB speech waveform concatenated with the upsampled NB one along the channel dimension. Each of the convolutional layers operates on kernels of size 32 with strides of 2. It has 6 layers with 16 channels in the first two layers, 32 channels in next two layers and 64 channels in the last two layers. No residual connections have been added and biases for all layers have been omitted. For activation, we use Leaky ReLU with negative slope of 0.2.

B. Training Objective

The adversarial metric used in this work is the hinge loss [29]:

$$L_{hinge} = max(0, 1 - tgt \cdot D()), \tag{1}$$

where $D()$ is the the raw output of the discriminator and tgt is the intended output: 1 for real speech and -1 for generated speech. Lim et al. [29] showed that hinge loss has less mode collapse and a more stable training behaviour compared to the loss used in the initial GAN paper [14] or the Wasserstein distance [30]. As already observed in [13], [15] the adversarial loss can be amended by an L^p-norm calculated on samples and on features. Here we use the L^1-norm calculated on time-domain samples and as feature loss L_{mel} the L^2-norm calculated on logarithmic Mel energies. The total loss training the generator is:

$$L = (1 - \lambda)L_{hinge} + \lambda(L^1 + L_{mel}). \tag{2}$$

IV. Experimental Setup

As training material we used several publicly available speech databases [31] as well as other speech items of different languages. In total, 13 hours of training material were used, all of it resampled to 16 kHz sampling frequency. Silent passages in the training data were removed with a voice-activation-detection [32]. The NB input signal was coded with AMR-NB at 10.2 kbps. All speech signals were pre-emphasised with a first order filter $E(z) = 1 - 0.68z^{-1}$ before entering the processing chain and the generated speech was inverse (de-emphasis) filtered with $E^{-1}(z) = \frac{1}{1-0.68z^{-1}}$. The reason for this is to compensate the spectral tilt of speech which may result in less pronounced high frequencies in the generated speech. The LPC envelope of order 12 is extracted on frames of 128 samples windowed with a Hann window by calculating the time-domain autocorrelation followed by the Levinson recursion. Thereafter they are converted to an FIR filter as explained in Sec. III-A. For the feature loss L_{mel}, 32 Mel energies are calculated on frames of 256 Hann-windowed samples with 50% overlap.

Both, LPC-GAN and CNN-GAN have 20 convolutional layers wither kernel-size 17 and 32 channels. LPC-GAN groups the channels into 4 groups, which has a huge impact on the computational complexity (see Tab.I).

The DNNs are trained with batches of 32 items with each item containing 1 second of speech. The optimisation algorithm for both the generator and discriminator is Adam [33] with a generator learning rate of 0.0001 and a discriminator learning rate of 0.0004. For a more stable adversarial loss, the coefficients used for computing running averages of the gradient and its square (the beta-parameters) are set to 0.5 and 0.99 respectively. The factor λ controlling the amount of feature-loss in Eq. 2 is set to 0.0015.

V. Evaluation

A. Computational Complexity

The computational complexity of the proposed SRs is an estimate of weighted million operations per speech-sample (WMOPS). WMOPS is the ITU unit for calculating computational complexity [34] of standardised speech processing tools. Additions (ADD), multiplications (MUL) as well as multiply-add (MAC) operations are each counted as one operation while complex operations like $tanh$, sigmoid or softmax operations each count as 25 operations. In the following sections, the number are calculated per speech-sample. This number is multiplied by the sampling frequency to get an estimate of the WMOPS. This should be seen as a rough approximation that does not consider advantages of todays parallel processing architectures. The results are summarised in Tab. I. Since the main field of application of SR is in speech coding, Tab. I also

TABLE I: Computational complexity and algorithmic delay of the proposed systems and EVS [35], [36], a state-of-the-art standardised speech codec. WMOPS is the ITU standard for calculating computational complexity [34] and calculated at a sampling frequency of 16kHz.

	OPS per sample	WMOPs	algorithmic delay
CNN-GAN:	1 387 897	22 206	22 ms
LPC-GAN:	383 353	6133	22 ms
EVS:	-	88	32 ms

contains the computational complexity of EVS [35], [36], the state-of-the-art standardised speech codec.

B. Algorithmic Delay

The algorithmic delay is the theoretical delay in ms between the input speech and the processed output speech caused by block-processing of speech samples. CPU or GPU time are not considered. The numbers are summarised in Tab. I. The source of algorithmic delay of the DNNs are the convolutional operations with kernels of size K. Each convolutional layer adds an algorithmic delay of $\lfloor K/2 \rfloor$ samples, since $\lfloor K/2 \rfloor - 1$ tabs of the kernel are calculated on previous samples and do not contribute to the delay. The algorithmic delay of the LPC processing is independent from the convolutional layer and can be neglected since the delay from the convolutional layers is always larger.

C. Objective Perceptual Quality

The evaluation of the quality of speech generated by GANs is a difficult task. In the typical use case GANs generate items from noise; metrics based on an L^p-norm cannot be used since there is no reference to compare with. In the following, we give state-of-the-art objective quality measures to see if they are able to predict the subjective ratings.

1) Perceptual Objective Listening Quality Analysis: Perceptual Objective Listening Quality Analysis (POLQA) is a standardised method that aims to predict the perceptual quality of coded speech signals on the same Mean Opinion Scale (MOS) used in listening tests [37]. First, masking thresholds are computed and then different kinds of distortions that exceed the masking threshold are calculated. These distortions are mapped to the MOS scale by a neuronal network. The estimated results are summarized in Fig. 4.

2) Fréchet Deep Speech Distance (FDSD): Since DNNs trained for recognition tasks are already quite elaborated, their output may be used for quality estimation. The Fréchet Deep Speech Distance (FDSD) proposed by Binkowski et al. [38] uses the raw output of *DeepSpeech 2* speech recognition network [39] to predict the quality of items generated by GANs. Fig. 4 gives the FDSD scores of the different SRs.

3) Word Error Rate (WER): Besides improving the perceptual quality, a SR can also improve the intelligibility of speech [2] and furthermore, the performance of Automatic Speech Recognition (ASR) systems. State of the art ASR systems are based on DNNs trained on uncompressed 16-kHz speech signals. As a result the performance of such systems drops significantly when the speech is coded with a NB codec. Fig. 4 depicts the impact on ASR of coding speech with AMR-NB. Postprocessing with SRs increases the ASR performance. The ASR system used here is Mozillas open implementation of *DeepSpeech* system [40].

WER is the only objective measure able to predict the human ratings (see next section).

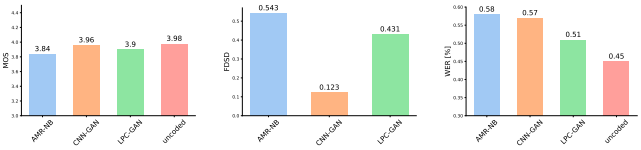

Fig. 4: Objective quality measures: (left) POLQA - higher values mean better quality; (middle) FDSD; (right) WER - lower values mean better quality for FDSD and WER.

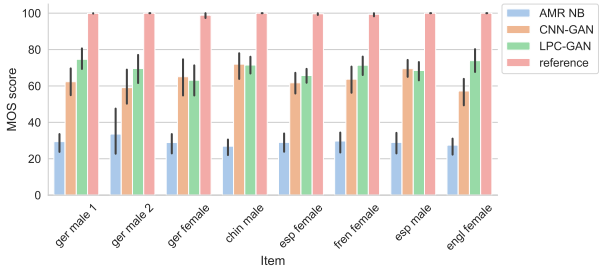

Fig. 5: Results from listening test evaluating different SRs as bar plot with 95% confidence intervals per item.

D. Subjective Perceptual Quality

To ultimately judge the perceptual quality of the proposed systems, a MUSHRA listening test [41] was conducted. According to the MUSHRA methodology, the test items contain the reference marked as such, a hidden reference and the AMR-NB coded signal serving as low anchor. 12 experienced listeners participated in the test. The speech items used in the test are about 10 seconds long and neither part of the training nor the test set. The items contain Chinese, English, French, German and Spanish speech from native speakers. The results are presented in Fig. 5. The results show that the proposed systems significantly improve the quality of AMR-NB speech for all items. None of the presented systems is significantly better than the others. The tendentially best system is the LPC-GAN.

VI. Conclusion

This work presents a novel approach for super-resolution of speech signals applying an established paradigm from the speech coding world, namely the decomposition of the speech signal into envelope and excitation signal (a.k.a. the source-filter model) to GANs. By conducting a listening test we could show that the perceptual quality is increased while lowering the computational complexity by a factor of more than 3. The proposed systems is also able to significantly improve the perceptual quality of AMR-NB coded speech. Furthermore, it improves the speech recognition WER of AMR-NB coded speech.

REFERENCES

[1] 3GPP, "TS 26.090, Mandatory Speech Codec speech processing functions; Adaptive Multi-Rate (AMR) speech codec; Transcoding functions," 1999.

[2] P. Bauer et al., "On improving telephone speech intelligibility for hearing impaired persons," in *Proceedings of the 10. ITG Conference on Speech Communication*, 2012.

[3] J. Abel et al., "A subjective listening test of six different artificial bandwidth extension approaches in english, chinese, german, and korean," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2016.

[4] P. Jax and P. Vary, "Wideband extension of telephone speech using a hidden markov model," in *IEEE Workshop on Speech Coding. Proceedings.*, 2000.

[5] K. Schmidt and B. Edler, "Blind bandwidth extension based on convolutional and recurrent deep neural networks," in *International Conference on Acoustics, Speech and Signal Processing*, 2018.

[6] K. Schmidt and B. Edler, "Deep neural network based guided speech bandwidth extension," in *Audio Engineering Society Convention 147*, Oct 2019.

[7] Kehuang Li et al., "A deep neural network approach to speech bandwidth expansion," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2015.

[8] J. Abel et al., "Artificial speech bandwidth extension using deep neural networks for wideband spectral envelope estimation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2018.

[9] Z. Ling et al., "Waveform modeling and generation using hierarchical recurrent neural networks for speech bandwidth extension," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2018.

[10] J. Sautter et al., "Artificial bandwidth extension using a conditional generative adversarial network with discriminative training," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2019.

[11] Yu Gu et al., "Speech bandwidth extension using bottleneck features and deep recurrent neural networks," in *Annual Conference of the International Speech Communication Association*, 2016.

[12] S. Li et al., "Speech bandwidth extension using generative adversarial networks," in *International Conference on Acoustics, Speech and Signal Processing*, 2018.

[13] S. E. Eskimez et al., "Adversarial training for speech super-resolution," *IEEE Journal of Selected Topics in Signal Processing*, 2019.

[14] I. J. Goodfellow et al., "Generative adversarial networks," 2014.

[15] S. Kim et. al, "Bandwidth extension on raw audio via generative adversarial networks," 2019.

[16] Y. Dong et al., "A time-frequency network with channel attention and non-local modules for artificial bandwidth extension," in *International Conference on Acoustics, Speech and Signal Processing*, 2020.

[17] X. Hao et al., "Time-domain neural network approach for speech bandwidth extension," in *International Conference on Acoustics, Speech and Signal Processing*, 2020.

[18] J. Su et al., "Bandwidth extension is all you need," in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2021, Toronto, ON, Canada, June 6-11, 2021*, 2021.

[19] K. Schmidt and B. Edler, "Blind bandwidth extension of speech based on lpcnet," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2020.

[20] Y. Tianren et al., "The computation of line spectral frequency using the second chebyshev polynomials," in *International Conference on Signal Processing*, 2002, vol. 1.

[21] S. Hochreiter et al., "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[22] L. Rabiner and R. Schafer, *Digital Processing of Speech Signals*, Englewood Cliffs: Prentice Hall, 1978.

[23] A. Mustafa et al., "Analysis by Adversarial Synthesis - A Novel Approach for Speech Vocoding," in *Proc. Interspeech*, 2019.

[24] K. He et al., "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016.

[25] A. Oord et al., "Wavenet: A generative model for raw audio," in *ISCA Speech Synthesis Workshop*, 2016.

[26] T. Salimans and D. P. Kingma, "Weight normalization: A simple reparameterization to accelerate training of deep neural networks," in *Advances in NeurIPS*, 2016.

[27] T. Miyato et al., "Spectral normalization for generative adversarial networks," 2018.

[28] M. Mirza and Simon Osindero, "Conditional generative adversarial nets," *ArXiv*, vol. abs/1411.1784, 2014.

[29] Jae Hyun Lim and Jong Chul Ye, "Geometric gan," 2017.

[30] M. Arjovsky et al., "Wasserstein gan," 2017.

[31] C. Veaux et al., "Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit," 2017.

[32] "Webrtc vad v2.0.10," https://webrtc.org.

[33] D. P. Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *CoRR*, 2014.

[34] ITU-T Study Group 12, *Software tools for speech and audio coding standardization*, Geneva, 2005.

[35] S. Bruhn et al., "Standardization of the new 3GPP EVS codec," in *International Conference on Acoustics, Speech and Signal Processing*, 2015.

[36] 3GPP TS 26.445, "EVS codec; detailed algorithmic description; technical specification, release 12," Sep. 2014.

[37] ITU-T Study Group 12, *P.863 : Perceptual objective listening quality prediction*, Geneva, 2018.

[38] M. Binkowski et al., "High fidelity speech synthesis with adversarial networks," *CoRR*, 2019.

[39] D. Amodei et al., "Deep speech 2: End-to-end speech recognition in english and mandarin," *CoRR*, 2015.

[40] Awni Y. Hannun et al., "Deep speech: Scaling up end-to-end speech recognition," *CoRR*, 2014.

[41] ITU-R, *Recommendation BS.1534-1 Method for subjective assessment of intermediate sound quality (MUSHRA)*, Geneva, 2003.