Artificial Bandwidth Extension of Narrowband Speech Using Generative Adversarial Networks

Jonas Sautter[1,2], Friedrich Faubel[1], Markus Buck[1], Gerhard Schmidt[2]

[1] Nuance Communications, Acoustic Speech Enhancement, Site Ulm, Germany, {jonas.sautter, friedrich.faubel, markus.buck}@nuance.com

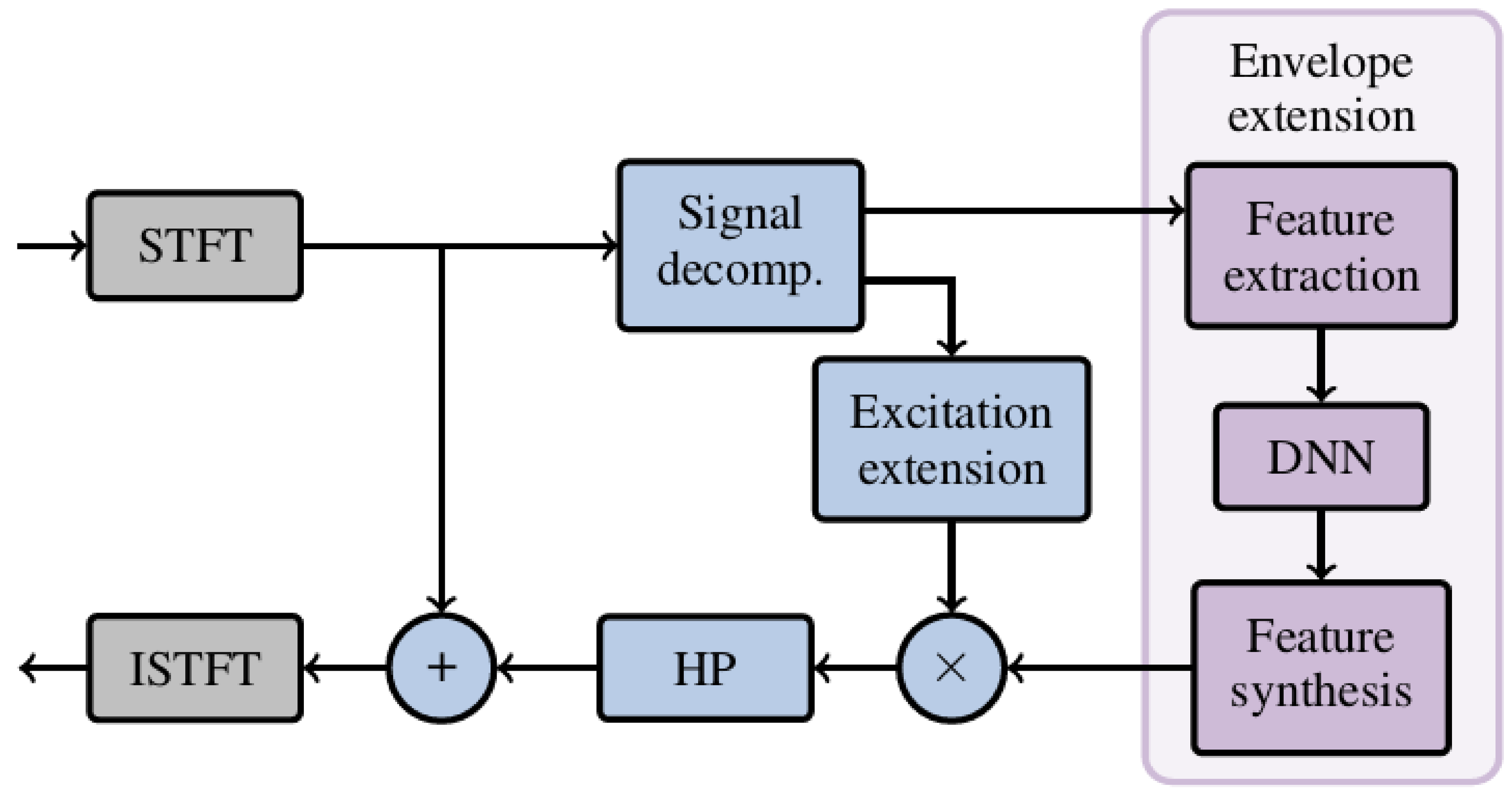
[2] Kiel University, Digital Signal Processing and System Theory, Kiel, Germany, gus@tf.uni-kiel.de

Summary

This work presents an artificial bandwidth extension (BWE) approach that restores high quality wideband speech from a low quality 4 kHz telephone signal. It uses a generative adversarial network (GAN) in combination with a discriminative cost function that better preserves the differences between fricatives and vowels. The combined approach gives an improvement of 1.7 in comparative mean opinion score (CMOS) over narrowband speech and an improvement of 0.8 over a standard GAN model.

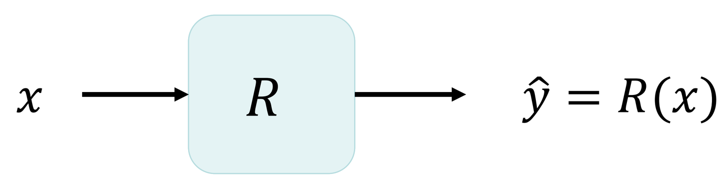
Artificial Bandwidth Extension

For the DNN-based BWE system, the STFT of the speech signal is decomposed into its spectral envelope and excitation. While the excitation is extended with traditional DSP techniques, envelope extension is performed with a regression DNN that has been trained to estimate 8 kHz wideband MFCCs from 4 kHz narrowband MFCC features.



MSE Training (DNN)

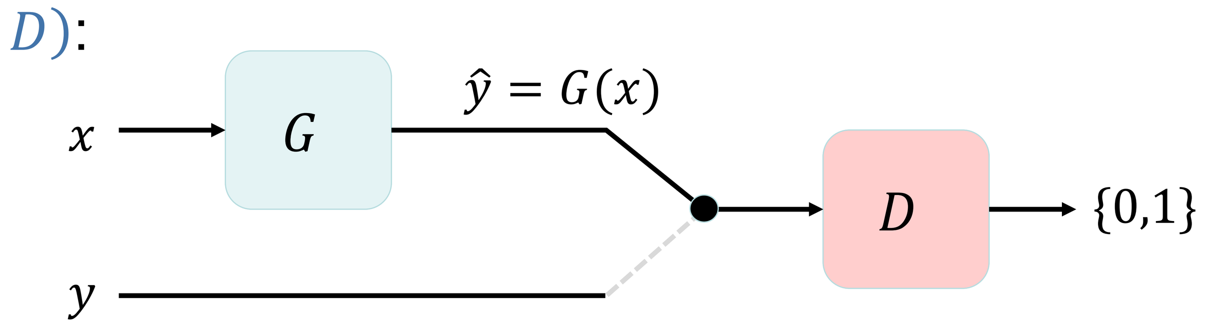
The baseline approach for BWE uses a simple regression DNN R. The mean squared error (MSE) between real and estimated wideband features, y and $\hat{y} = R(x)$, is used as loss function that is to be minimized during training:



$$\mathcal{L}_{MSE}(R) = \mathbb{E}_{x,y}[\|y - R(x)\|^2]$$

Generative Adversarial Network (GAN) Training

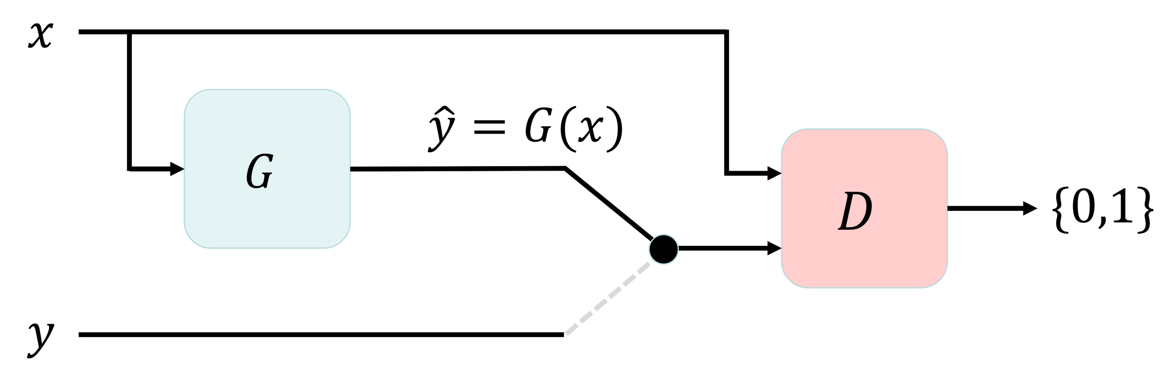
For GAN training, the regression network R is replaced by a generator network G, and an auxiliary discriminator network D is added. These networks have opposite tasks. While the generator network learns to generate wide-band spectra $\hat{y}$ that can hardly be distinguished from real spectra y, the discriminator network learns to distinguish y from $\hat{y}$. This is achieved by alternatingly training D to maximize $\mathcal{L}_{GAN}(G, D)$ and training G to minimize $\mathcal{L}_{GAN}(G, D)$:



$$\mathcal{L}_{GAN}(G, D) = \mathbb{E}_y[\log(D(y))] + \mathbb{E}_x[\log(1 - D(G(x))]$$

Conditional GAN (CGAN) Training

In conditional GANs, the input x of the generator is additionally given to the discriminator. This enables the discriminator to judge if the wideband spectrum it receives is real, conditioned on the narrowband input x.



Discriminative Loss Term

While GAN training reduces the amount of over-smoothing that is occasionally seen in bandwidth-extended spectra, it does not completely resolve the problem. Hence, we add a discriminative term to the loss function that explicitly preserves the upper band (4 to 8 kHz) power ratio between sharp fricatives ("s", "sh", ...) and other phonemes, here called $SFPR$:

$$\mathcal{L}_{DISC}(G) = \left|\frac{SFPR(G(x)) - SFPR(y)}{SFPR(y)}\right|^2$$

Proposed Training Objective (CGAND)

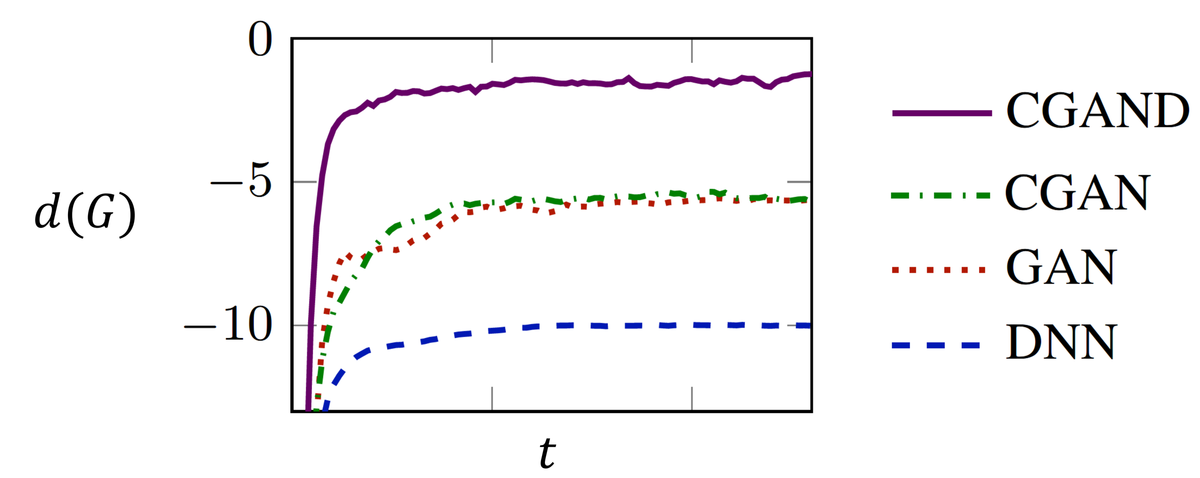
In the proposed system, we combine all of the above loss terms in one objective that is to be optimized during training:

$$\min_G \max_D \ \mathcal{L}_{MSE}(G) + \lambda_{GAN} \cdot \mathcal{L}_{GAN}(G, D) + \lambda_{DISC} \cdot \mathcal{L}_{DISC}(G)$$

Experimental Results

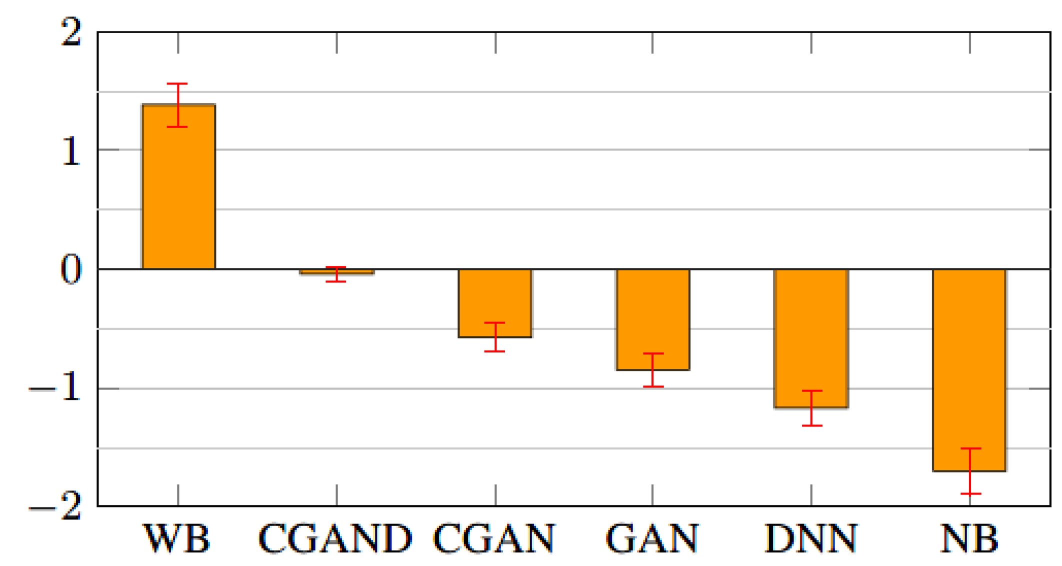
Objective Evaluation

The effect of over-smoothing can be observed in the variance of the upper band energy per mini-batch $\sigma_{\mathrm{UB}}{}^2$. This variance is too low after processing with the basic DNN. Training a GAN instead and using the discriminative loss term decreases this mismatch. The deviation $d(G) = \sigma_{\mathrm{UB}}(G(x)) - \sigma_{\mathrm{UB}}(y)$ is shown in the following figure for the whole training process:



Subjective Evaluation

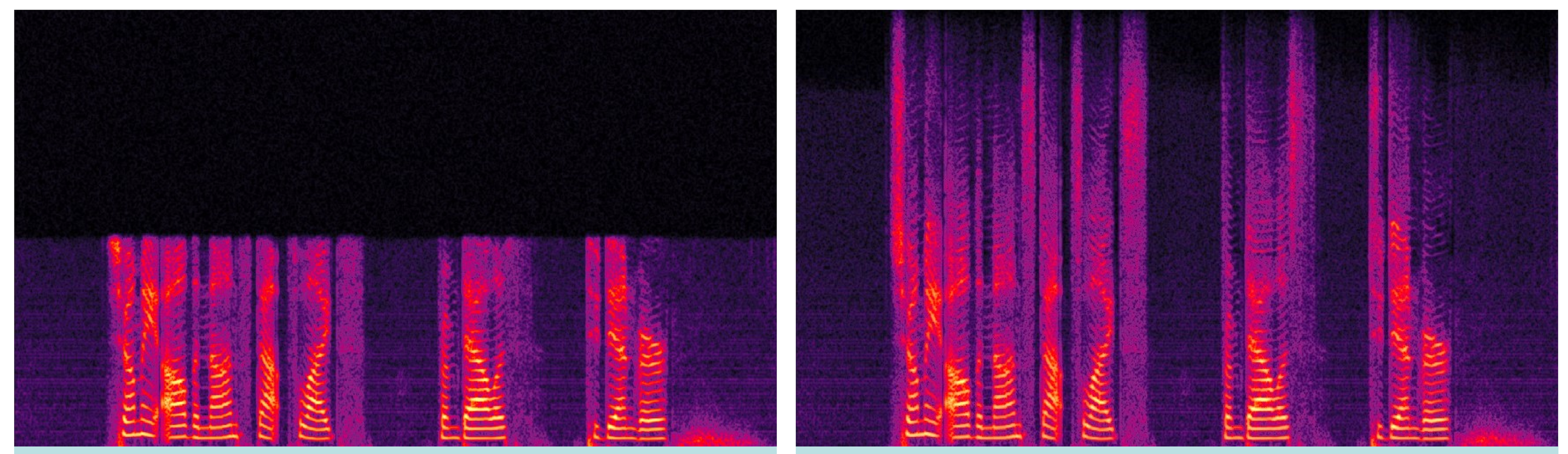
The following diagram shows comparative mean opinion scores (human rating between -3 and 3) of the proposed CGAND system compared to the real wideband signal (WB), the original narrowband signal (NB), a base-line system with MSE training (DNN), a GAN as well as a CGAN. The error bars (red lines) indicate 95% confidence intervals.



Comparative Mean Opinion (CMOS) Scores [-3, 3]

Conclusion:
The proposed system significantly outperforms other BWE methods

Example Spectrogram



4 kHz narrowband signal | 8 kHz bandwidth extended signal

International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2019)
Brighton, UK, May 12–17, 2019

Digital Signal Processing and System Theory

Christian-Albrechts-Universität zu Kiel