Moliner, Eloi; Valimaki, Vesa

BEHM-GAN: Bandwidth Extension of Historical Music using Generative Adversarial Networks

Published in:
IEEE/ACM Transactions on Audio Speech and Language Processing

DOI:
10.1109/TASLP.2022.3190726

Published: 01/01/2023

Document Version
Publisher's PDF, also known as Version of record

Published under the following license:
CC BY

Please cite the original version:
Moliner, E., & Valimaki, V. (2023). BEHM-GAN: Bandwidth Extension of Historical Music using Generative Adversarial Networks. *IEEE/ACM Transactions on Audio Speech and Language Processing*, *31*, 943-956. https://doi.org/10.1109/TASLP.2022.3190726

BEHM-GAN: Bandwidth Extension of Historical Music Using Generative Adversarial Networks

Eloi Moliner and Vesa Välimäki, *Fellow, IEEE*

***Abstract*—Audio bandwidth extension aims to expand the spectrum of bandlimited audio signals. Although this topic has been broadly studied during recent years, the particular problem of extending the bandwidth of historical music recordings remains an open challenge. This paper proposes a method for the bandwidth extension of historical music using generative adversarial networks (BEHM-GAN) as a practical solution to this problem. The proposed method works with the complex spectrogram representation of audio and, thanks to a dedicated regularization strategy, can effectively extend the bandwidth of out-of-distribution real historical recordings. The BEHM-GAN is designed to be applied as a second step after denoising the recording to suppress any additive disturbances, such as clicks and background noise. We train and evaluate the method using solo piano classical music. The proposed method outperforms the compared baselines in both objective and subjective experiments. The results of a formal blind listening test show that BEHM-GAN significantly increases the perceptual sound quality in early-20th-century gramophone recordings. For several items, there is a substantial improvement in the mean opinion score after enhancing historical recordings with the proposed bandwidth-extension algorithm. This study represents a relevant step toward data-driven music restoration in real-world scenarios.**

***Index Terms*—Audio recording, convolutional neural networks, machine learning, music, signal restoration.**

I. Introduction

HISTORICAL music recordings are available in large numbers in archives but, due to the technological limitations of the time, by modern standard they are of a very poor audio quality. Early-20th-century gramophone recordings suffer from severe degradations, such as multiple kinds of surface noises, distortion, and a narrow frequency bandwidth [1], [2]. The goal of digital audio restoration is to correct the imperfections of audio recordings so that the resulting sound quality is enhanced. Restoration may target the removal of clicks and noises [3], [4], the inpainting of missing audio segments [5], [6], declipping [7], or the bandwidth extension of bandlimited audio signals, among other tasks.

Manuscript received 16 February 2022; revised 15 June 2022; accepted 17 June 2022. Date of publication 14 July 2022; date of current version 15 February 2023. This work was supported by the activities of the Nordic Sound and Music Computing Network—NordicSMC, NordForsk Project 86892. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Timo Gerkmann.*(Corresponding author: Eloi Moliner.)*

The authors are with the Acoustics Laboratory, Department of Information and Communications Engineering, Aalto University, 02150 Espoo, Finland (e-mail: eloi.moliner@aalto.fi; vesa.valimaki@aalto.fi).

Digital Object Identifier 10.1109/TASLP.2022.3190726

This paper focuses on audio bandwidth extension and, particularly, on applying it to historical music recordings. During recent years, many works have used modern deep learning technologies for bandwidth extension. However, their goal has usually been to enhance modern digital audio signals having a limited bandwidth because of the usage of a lower sampling rate. Only a few exceptions are relevant to music signal processing [8], [9], [10], whereas most of these studies focus on processing speech [11], [12], [13], [14], [15]. Although music and speech share the same domain of acoustic signals, the two are fundamentally different.

Usually, the aforementioned methods are trained in a self-supervised fashion by pre-processing the audio data with lowpass filters to simulate the bandwidth limitation. Then, the models are optimized to extend the input lowpass-filtered audio using the broadband original signal as a target. However, bandwidth-extending historical recordings entails an extra challenge, as no full-bandwidth version is available for this particular material. Then, we would rely on the model, trained with synthetically filtered data, to extrapolate to real historical recordings, a harder out-of-distribution scenario. One should also consider the problem of filter generalization [16], which refers to the inability deep neural networks to generalize when they are trained using a single type of lowpass filter in the training-data pipeline.

Another problem with old gramophone recordings is that they are often corrupted with a wide range of global and local disturbances, such as hiss, clicks, and thumps. These additive noises represent another obstacle in enhancing the recording. Luckily, recent works have shown that a vast majority of clicks and noises appearing on gramophone recordings can be efficiently suppressed using deep-learning models [4], [17]. We studied this problem in particular and proposed a model consisting of a spectrogram-based deep-neural-network architecture [17]. The denoising model was trained using a realistic dataset of noise samples extracted from gramophone recordings and yields a considerable enhancement in quality [17]. This paper builds upon this previous work [17], in such a way that the proposed bandwidth extension method is intended to be applied as a second step after the original recording has been first denoised, as illustrated in Fig. 1.

In this paper, we present a method for the bandwidth extension of historical music recordings using generative adversarial networks (BEHM-GAN) and evaluate it with solo piano music recordings. The proposed method is based on a generative adversarial network (GAN) [18] and combines a generator in the spectrogram domain with multiple time-domain discriminators. To provide the model with the necessary robustness to make

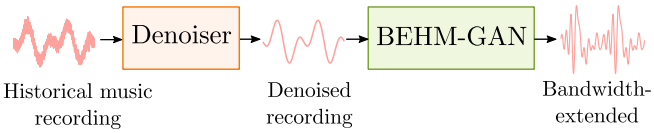

Fig. 1. Illustration of the inference pipeline. The denoiser block refers to a model borrowed from earlier work [17], and the BEHM-GAN is the model proposed in this paper.

inference in historical recordings, we propose a simple but effective filter regularization in the training stage. The strategy is based on adding a small amount of white Gaussian noise after applying a lowpass filter with a randomized cutoff frequency. We show that the proposed method inserts sound energy in an appropriate way above the cutoff frequency of about 3 kHz and, as determined in a formal blind listening test, significantly improves the perceived quality of both artificially bandlimited and real old piano music. As far as we are aware, this is the first work that successfully extends the bandwidth in real historical music recordings. We emphasize that the goal of this work is not to restore exactly the missing sound events, but to recreate plausible high-frequency content and, thus, make the music more pleasant to listen to.

The remainder of this paper is organized as follows. Section II reviews the most relevant related work, with a focus on recent deep-learning-based audio-bandwidth-extension studies. To understand the bandwidth limitation of historical gramophone recordings, Section III analyzes empirically their spectral characteristics using the Long-Term Average Spectrum (LTAS). The proposed BEHM-GAN method is introduced in Section IV and is evaluated in Section V with objective and subjective metrics. In order to assess the robustness of the compared methods, the experiments are conducted under two separate conditions: lowpass filtered modern recordings and old historical recordings. Finally, Section VI concludes the paper.

II. Related Work

This section reviews previous work on audio bandwidth extension, focusing on approaches that apply GANs for related tasks or study the problem of filter generalization.

A. Audio Bandwidth Extension

Audio bandwidth extension refers to methods that extend the spectrum of audio signals [19]. A popular sub-topic is audio super-resolution [11], [12], which increases the sampling rate of a given audio signal by extending its bandwidth above the original Nyquist limit. This topic has a long history in telephony, where the bandwidth of a transmitted speech signal was usually compressed because of channel constraints [20]. Another relevant application is audio compression, as bandwidth-extension techniques can be used to reduce the bit rate of an audio signal [21], [22].

Early works used signal processing methods such as a source-filter model [23], [24], codebook mapping [25], nonlinear devices [19], or spectral band replication [26]. Other approaches were based on data-driven techniques, such as Gaussian mixture models [27], [28], hidden Markov models [29], or shallow neural networks [30], [31]. However, due to their inadequate modeling capabilities, these early methods often lead to a poor or mediocre audio quality.

More recently, deep-learning-based bandwidth extension-methods outperformed previous approaches. The vast majority of the presented methods used convolutional neural networks and work using either a spectrogram representation [32], [33], raw audio data [11], [12], [34], [35], or a mixture of both [36], [37].

B. GANs for Audio Bandwidth Extension

Deep learning models based on optimizing reconstruction losses excel at tasks where the goal is to design a nonlinear mapping between two data distributions, e.g., denoising. However, the performance of supervised learning is limited when the task involves generating new content that is absent in the observed signal, as is the case in bandwidth extension. As a consequence, deep learning tends to build over-smoothed and unrealistic spectra. For this reason, recent works have adopted a generative approach that allows the model to have more expressive power. Some studies applied different kinds of generative models for the task of speech bandwidth extension, such as flow-based models [38] or diffusion-probabilistic models [15], but, in particular, GANs [18] have shown great potential for this task.

GANs are generative models that are based on optimizing a two-player min-max game between a generator G and a discriminator D [18]. The discriminator D is optimized to distinguish real data samples from the ones generated by G, whereas G tries to fool D by generating data samples that are harder to detect. Ideally, if the training does not collapse, both G and D will converge to the so-called Nash equilibrium, where G fits the target data distribution and D is unable to detect the fake data samples from the real ones. In the original GAN formulation, a latent vector of Gaussian noise z is provided to the generator $G(z)$ as an input. However, GANs for audio bandwidth extension can be viewed as conditional GANs [39], where the generator $G(x, z)$ is also conditioned on an observed signal x, here the bandlimited input. Then, due to the high dimensionality of x, the latent vector z is often omitted if a controllable latent-space representation is not required [40].

Although only a few studies have applied GAN models to bandwidth extension of music signals [10], [41], many recent works have applied them for speech [13], [14], [42]. Eskimez et al. [42] proposed one of the earliest works using an adversarial approach for speech super-resolution. Their proposed model predicted the magnitude spectrogram representation of audio and used an adversarial loss combined with a reconstruction loss. However, the Eskimez model had the limitation that it did not predict the phase information but just replicated it [42]. Other phase-aware works made an effort to incorporate the phase information into the training framework [10]. Instead, Kim et al. [41], opted for working directly on raw audio, thus avoiding the aforementioned phase issues. They also incorporated a third auxiliary feature-matching loss term. Su et al. [13] used a

time-domain Wavenet generator and a composite of multiple time-domain and spectral-domain discriminators. Utilizing a complex combination of loss terms, they achieved impressive results. Li et al. [14] proposed a lighter time-domain model that was suitable to run in real-time.

C. Lowpass Filter Generalization

A particular problem in the audio-bandwidth-extension literature is the incapability of deep neural networks to generalize when they are trained using lowpass filters. We hypothesize that this problem is a special case of shortcut learning [43], stating that the model does not learn the true underlying mechanisms of the data but relies on spurious statistical relationships. In this case, the neural network learns the easier task of inverting the response of a lowpass filter instead of generating new and coherent high-frequency content.

Kuleshov et al. [11] observed that a neural network trained to conduct audio super-resolution using aliased training data was ineffective if an antialiasing filter was included during testing. The same problem happened when antialising filters were only used during training and when the filters utilized during training and testing differed. Sulun and Davies [16] studied this phenomenon and named it "filter overfitting". They showed that the problem could be mitigated considerably by using a set of different lowpass filters during training as a data augmentation strategy.

Wang and Wang [12] examined the robustness of their speech super-resolution model that was trained with different down-sampling schemes. They proposed a solution based on randomly combining three different down-sampling strategies during training. Li et al. [14] also experimented with using variable-band filters with randomized cutoff frequencies to increase the robustness of the model in real-life speech bandwidth-extension scenarios. Similarly, Nguyen et al. [44] applied anti-aliasing filters having random order and ripple intending to improve the robustness of their model.

This problem gets more relevant in the case of historical recordings when we aim to infer a target distribution that has not been processed by any lowpass filter. In this case, neither a specific filter specification nor a known cutoff frequency can be assumed, since they may vary greatly depending on the recording conditions. In the next section, we investigate the underlying lowpass filtering in old gramophone recordings.

III. Spectral Analysis of Gramophone Recordings

To obtain prior knowledge to design our method, we analyzed the bandwidth of 78-RPM (rounds per minute) gramophone recordings, which we were interested in enhancing. To fully understand the frequency characteristics of these recordings, one must study the recording conditions of the time. However, due to the lack of international standards, the exact characteristics vary widely depending on the manufacturer, the publication date, the recording material, or possible equalization corrections made by recording engineers. Hence, the work of audio restoration is extremely hard, as restoration engineers now have to conduct a study on industrial archaeology for every single record they aim to restore [45].

One of the main reasons for the limited bandwidth in old analog recordings are the disc-cutting lathes, used to record sound into the physical disc media [45]. The most critical piece, the cutterhead, converts electric waveforms into modulations in a groove. The frequency response of the recording vary greatly depending on the cutterhead model, the speed of the record, or the shape of the stylus. The most commonly used cutterheads during the early 1920 s produced a resonance frequency between 3 kHz and 4 kHz, and above that the frequency response decayed rapidly. As a consequence, due to the poor signal-to-noise ratio (SNR) of the recordings, the high-frequency components above this resonance frequency were practically lost. With the introduction of Western Electric's electromechanical cutterhead in 1925, the frequency response could be considerably flattened, but the cutoff frequency could not be extended to above 5 kHz [45]. Over the years, better equipment was developed that allowed engineers to extend the recordable bandwidth of audio, thanks to many technological advances like motional feedback [45].

To analyze empirically the spectral characteristics of 78-RPM gramophone recordings, we conduct a study based on the LTAS. To do so, we collected six 78-RPM gramophone recordings of a given music piece, all of them containing a similar ensemble of instruments and dated from 1920 to 1930. For comparison, we also collected three contemporary broadband recordings of the same piece. The six old recordings were first denoised using our previously proposed method [17]. The LTAS was calculated for each of the recordings using the IoSR library [46], applying Gaussian smoothing per octave band. We then subtract the LTAS of the three contemporary recordings from each of the old versions to obtain a rough estimate of the frequency response of the recording. The resulting 18 difference LTAS curves are re-scaled so that their mean level between 500 Hz and 2 kHz is 0 dB.

Fig. 2 shows the computed difference LTAS curves and their average for three classical pieces: the orchestral piece *The Blue Danube Waltz*, by Johann Strauss (Fig. 2(a)); the opera piece *L'amour est un oiseau rebelle*, from Carmen by Georges Bizet (Fig. 2(b)); and *Humoresque No. 7*, by Antonin Dvořák, played by string ensembles (Fig. 2(c)). Although these plots do not give accurate information due to the averaging and the octave-band smoothing, they indicate a decaying trend starting at approximately 3 kHz. The estimated -3-dB cutoff frequencies are 2.7, 3.1, and 2.5 kHz for the above-mentioned historical recordings, as indicated in Fig. 2.

IV. Behm-Gan

This section presents the BEHM-GAN, the proposed GAN-based method for the bandwidth extension of historical recordings. The generator model, the loss functions, and the three different discriminators used are first described, and their roles are also illustrated in Fig. 3. The dataset, the use of lowpass filters to simulate the loss of high frequencies, and the implementation of the training are also explained.

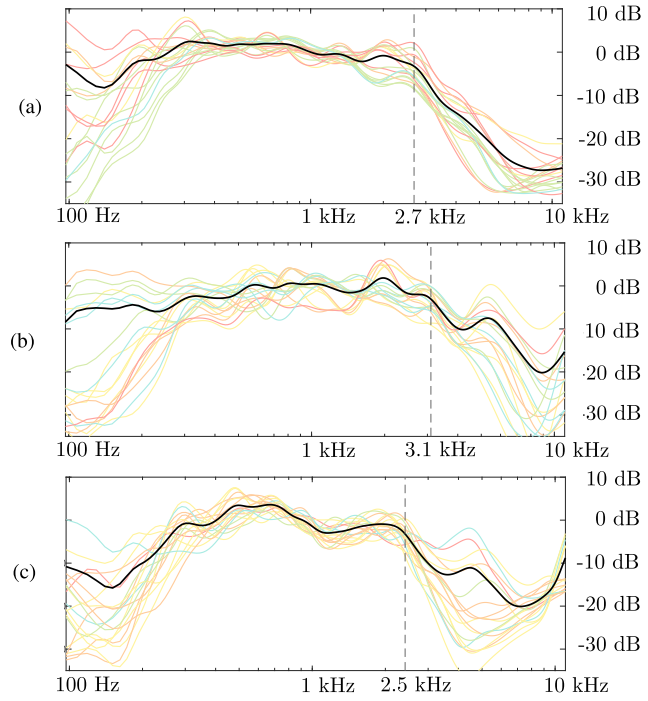

Fig. 2. LTAS difference curves computed between six pre-1930s recordings and three contemporary recordings of (a) *The Blue Danube Waltz*, (b) *Carmen*, and (c) *Humoresque*. Each colored line represents the difference between the LTAS of one of the six old recordings and one of the three modern ones, totalling 18 curves. The black line is the average of the difference curves, and the vertical dashed line marks its −3-dB point.

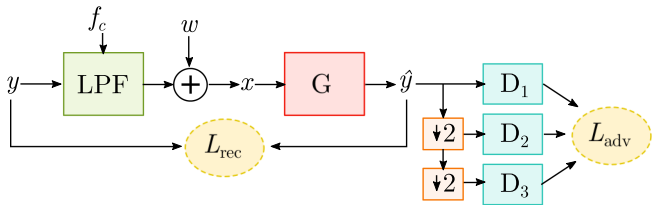

Fig. 3. Proposed GAN-based training framework containing generator G, three time-domain discriminators D_1, D_2, and D_3, a variable lowpass filter (LPF), and additive noise w. The training is optimized with a composite of two losses: an adversarial loss L_{adv} and a reconstruction loss L_{rec}. The down-sampling operators refer to strided average pooling with a kernel size of 4.

A. Generator Model Architecture

We opt to work in the time-frequency domain to design our generator. Thus, the input bandlimited audio x sampled at $f_{\text{s}} = 22.05$ kHz is first transformed by means of the short-time fourier transform (STFT). We use an FFT length of 1024 samples (46.11 ms) with a Hamming window of the same size and a hop length of 256 samples (11.61 ms). The resulting complex signal is converted to a real one by stacking its real and imaginary parts as separate channels. Then, assuming that the generator succeeds at maintaining the implicit phase information, the output signal can be directly converted back to the time domain using the inverse-STFT, without the need for a phase-recovery technique. We opted not to use a complex-aware neural network architecture [47] as, in our experiments, it did not provide any clear benefit against the double-real representation. Moreover, complex-valued modules require twice the amount of computation, an issue that slowed down the training significantly.

The generator architecture is based on the U-Net model [17] and is shown in Fig. 4. The architecture is formed by 2D-convolutions and Exponential Linear Unit non-linearities [48] to capture time-frequency features from the spectrogram. We concatenate frequency-positional embeddings [49] as an inductive bias to break the frequency-equivariance symmetry, which is implicit in 2D-convolutions. The architecture has an encoder-decoder structure with residual DenseNet blocks [50] as intermediate layers. The encoder coarsens the resolution at each layer using strided convolutions, sequentially increasing the number of channels. The decoder structure is symmetrical to the encoder and upsamples the resolution with transposed convolutions. The concatenative skip connections help to retain fine-grained details of the spectrogram. We refer the reader to the source code[1] for further details on the model implementation and the used hyperparameters.

B. Training Objective

The generator is optimized with a composite of two losses, an adversarial loss L_{adv} and an auxiliary reconstruction loss L_{rec}:

$$L_G = L_{\text{adv}} + \alpha L_{\text{rec}}, \tag{1}$$

where the coefficient $\alpha = 0.4$ is a tuning hyperparameter used to combine the two loss terms. The value of α was optimized by grid search, using informal listening as the quality criterion. For the adversarial loss, we adopt the multi-scale discriminators D_1, D_2, and D_3 from MelGan [51].

As indicated in Fig. 3, discriminator D_1 operates directly on the raw audio waveform, whereas the input waveforms of discriminators D_2 and D_3 are, respectively, downsampled by factors 2 and 4. Thus, each discriminator learns features in a different frequency range. Since the model operates at $f_{\text{s}} = 22.05$ kHz, D_1 observes frequency components up to the Nyquist limit $f_{\text{s}}/2 = 11.03$ kHz, D_2 up to $f_{\text{s}}/4 = 5.51$ kHz and D_3 only to $f_{\text{s}}/8 = 2.76$ kHz. Although our main interest is to reconstruct the frequency components above 3.0 kHz, using D_3 is still beneficial to stabilize the adversarial training and leads to better convergence. The architecture of each of the discriminators consists of a stack of grouped strided convolutions, and the down-sampling is performed by strided average pooling with a kernel size of 4, in the same way as in [51]. Time-domain discriminators are highly sensitive to phase mismatches in the data. This is a very convenient property when using a spectrogram-based generator, since maintaining the phase coherence when the audio data is transformed to the complex STFT domain may be problematic.

In this work, we apply the least-squares GAN objective [52]. The adversarial loss for the generator is then defined as

$$L_{\text{adv}} = \mathbb{E}_{\hat{y_k}}\left[\sum_k (D_k(\hat{y_k}) - 1)^2\right], \tag{2}$$

[1][Online]. Available: https://github.com/eloimoliner/bwe_historical_recordings

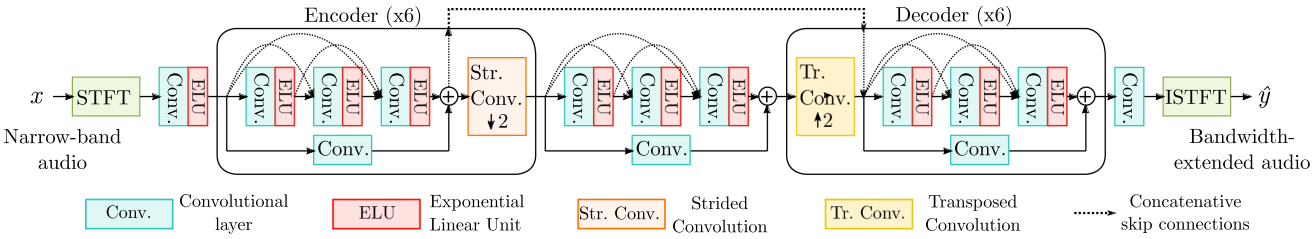

Fig. 4. Proposed U-net-based architecture of the generator model, cf. Fig. 3.

where $\mathbb{E}$ is the expectation operator. The discriminators are optimized by minimizing the loss function:

$$L_{D_k} = \frac{1}{2}\mathbb{E}_{y_k}\left[(D_k(y_k)-1)^2\right] + \frac{1}{2}\mathbb{E}_{\hat{y}_k}\left[D_k(\hat{y}_k)^2\right]. \quad (3)$$

We use the multi-resolution STFT loss [53] as the auxiliary reconstruction loss L_{rec}. This loss is defined as the expectation of the sum of two terms, $L_{\text{sc}}^{(m)}$ and $L_{\text{mag}}^{(m)}$, at M different frequency resolutions as

$$L_{\text{rec}}^{(m)} = \mathbb{E}_{y,\hat{y}}\left[\frac{1}{M}\sum_{m=1}^{M} L_{\text{sc}}^{(m)} + L_{\text{mag}}^{(m)}\right]. \quad (4)$$

The spectral convergence term $L_{\text{sc}}^{(m)}$ and the log magnitude distance term $L_{\text{mag}}^{(m)}$ are defined, respectively, as:

$$L_{\text{sc}}^{(m)} = \frac{\| \, |Y^{(m)}| - |\hat{Y}^{(m)}| \, \|_{\text{F}}}{\| \, |Y^{(m)}| \, \|_{\text{F}}} \quad (5)$$

and

$$L_{\text{mag}}^{(m)} = \frac{1}{S}\| \log|Y^{(m)}| - \log|\hat{Y}^{(m)}| \|_1, \quad (6)$$

where $Y^{(m)}$ and $\hat{Y}^{(m)}$ are the STFTs of the signals y and $\hat{y}$, respectively, using an analysis window of length $m \in \{256, 512, 1024, 2048\}$, $\|\cdot\|_{\text{F}}$ is the Frobenius norm, $\|\cdot\|_1$ is the L1 norm, and S is the total number of STFT bins. Note that this reconstruction loss is only aware of the magnitude differences in the spectrogram, as we rely on the adversarial loss term to deal with the phase information.

C. Dataset

We train and evaluate our method using solo piano classical music. Doing so, we reduce the difficulty of the problem by limiting the variance in the training data. Piano sounds are a convenient choice for evaluating bandwidth-extension algorithms, since they contain both transient and tonal components. Moreover, since the piano is one of the most common musical instruments for solo performances, a large quantity of contemporary and historical solo piano recordings is publicly available. Considering that classical music is a genre that has practically remained unchanged over time, we avoid introducing a major divergence between the training and target distributions.

We collected our training data from the solo piano pieces of the MusicNet dataset [54], but discarded some of the older recordings as they contained heavy background noise and the audio quality was suboptimal. The training set contains 14.4 h of broadband piano classical music. A separate test set with 1.1 h of broadband piano music is used for the objective and subjective evaluation metrics that require a reference signal (Section V-B and Section V-C). The music pieces included in the test set are not present in the training set.

A test set of real historical recordings was also collected to compute the objective and subjective evaluation metrics that do not require a reference signal (Section V-B and Section V-D). It consists of six historical solo piano recordings extracted from "The Great 78 Project" [55], a large collection of publicly available digitized 78-RPM gramophone records [56].

D. Lowpass Filter Generalization

Since the frequency responses in historical music recordings are far from being deterministic, we apply a lowpass filter with a randomized cutoff frequency f_{c} to the training data. We parameterize the cutoff frequency with a normal distribution, whose mean and standard deviation are set up empirically to be a rough estimate of the frequency responses in the gramophone recordings in the 1920 s.

Based on our findings from the spectral analysis in Section III, we set the mean cutoff frequency to $\mu_{f_{\text{c}}} = 3.0$ kHz and the standard deviation to $\sigma_{f_{\text{c}}} = 300$ Hz. The value of $\sigma_{f_{\text{c}}}$ is the result of a trade-off off bias against variance in the model. In other words, a model trained with a larger $\sigma_{f_{\text{c}}}$ would probably generalize to a wider range of cutoff frequencies. However, the resulting quality would likely diminish due to the increase in variance in the data, making the optimization more challenging and unstable. The lowpass filters are 25th-order FIR filters using the windowing method with the Kaiser window ($\beta = 1$). The magnitude responses of the FIR filters used for training are presented in Fig. 5. FIR filters have the convenient advantage that they can be efficiently implemented by applying convolution, thus not demanding much extra computation during training.

The idea of applying variable-band filters was also studied by Li et al. [14], with the difference that they used a uniform distribution instead of a normal one. Randomizing the cutoff frequency of the filter indeed helps to increase the robustness of the model in different frequency ranges, but, as we show with our experiments in Section V-D, randomization is certainly not enough if we want to successfully make inferences in historical recordings.

To further regularize the model, we apply a well known regularization approach [57] and corrupt the lowpass-filtered training data by adding a small amount of Gaussian white

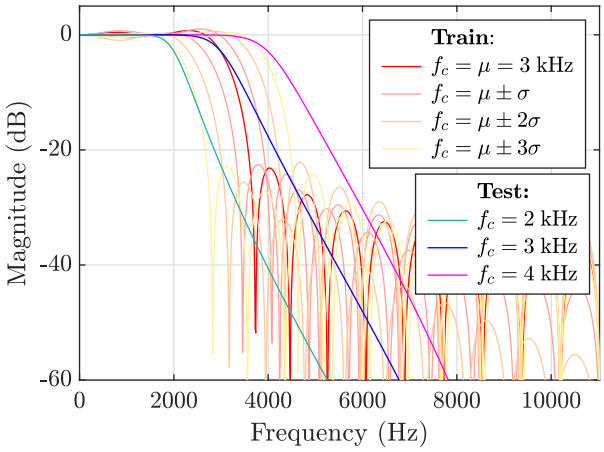

Fig. 5. Magnitude responses of lowpass filters used for training and testing. The training filters are FIR with a randomized cutoff frequency, while the testing filters are IIR with a fixed cutoff.

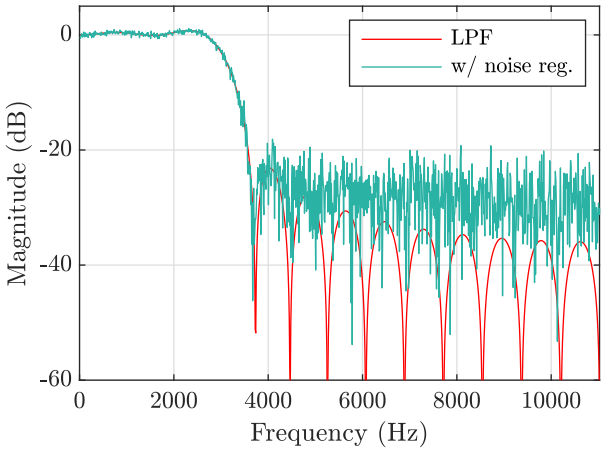

Fig. 6. Magnitude response of one of the lowpass filters used during training and its equivalent response after applying the noise regularization.

noise w having zero mean and a fixed power $\sigma^2 = -30$ dBFS (decibels relative to full scale) directly to the raw audio signal. The added noise diffuses the magnitude response of the filter, as Fig. 6 demonstrates, and successfully enforces the model to generate new high-frequency content instead of overfitting the filter shape. The use of the additive noise during training also encourages the generator to focus only on the most prominent musical features making the model robust to the minuscule denoising residuals that it may encounter while making inferences in historical recordings. Furthermore, this noise regularization strategy injects stochasticity into the model. Given that we do not add a latent vector z to the generator, this extra noise allows the generator to produce stochastic outputs. The fixed value of $\sigma^2 = -30$ dBFS is chosen in consistence with the training lowpass filters, in such a way that the noise level is sufficiently high to mask the remaining information in the side lobes of the filter (see Fig. 6). We observed that using higher values of σ^2 was often detrimental to the resulting audio quality, as some unwanted noisy residuals were still present in the output.

E. Making Inferences in Historical Recordings

Using the proposed regularization, the generator can be applied to make inferences on out-of-distribution historical music recordings. Our inference pipeline builds on previous work on denoising [17], as illustrated in Fig. 1. The original noisy recordings are first denoised to suppress clicks, hisses, and other additive disturbances. Then, the denoised recordings are bandwidth-extended by directly applying the pre-trained STFT-based generator.

In the same way as during training, the noise regularization could be added at the inference stage before feeding the denoised recording to the bandwidth-extension generator. As discussed in Section V-B, this step helps to achieve better objective metrics. However, in the majority of the tested cases, no perceptual differences were noticed with or without noise during inference and hence it is left as an optional step.

F. Implementation Details

The used sampling frequency $f_s = 22.050$ kHz sets the upper limit of processing to about 11 kHz. This choice makes the training fast and still leaves a wide range from about 3 to 11 kHz for bandwidth extension. For training, we used batches of four audio segments, each with a duration of 5 s. Nevertheless, due to the nature of convolutional neural networks, the input length can be set arbitrarily during inference. With the goal to make the model robust to different volume (loudness) levels, we also apply a uniformly random gain, set between -6 dB and 4 dB for each input signal. We did not find using batch normalization or weight normalization beneficial to the generator. The discriminators, however, are weight-normalized [58].

We use the Adam optimizer [59] with the parameters $\beta_1 = 0.5$ and $\beta_2 = 0.9$ to train both the generator G and the discriminators D_k. The training is divided into two separate stages. First, we train G for 10,000 steps with a learning rate of 1×10^{-4} using only the reconstruction loss L_{rec}. This step guarantees that the model learns to apply an identity mapping to the low-frequency components before including the adversarial discriminators into the training loop. Then, we decrease the learning rate to 1×10^{-5}, incorporate the adversarial loss L_{adv}, and continue training for 300,000 steps. During the second stage, the discriminators D_k are updated twice for every step taken by the generator, using a learning rate of 1×10^{-4}. The training took, on average, two days to complete on a single Tesla V100 GPU in Triton, Aalto University's computing cluster.

V. Experiments and Results

This section evaluates the quality of the bandwidth-extension using both objective and subjective experiments.

A. Comparison Models

We compare our proposed method with two baseline models, AudioUnet [11] and SEANet [14]. Considering that these baselines were not designed to do bandwidth extension in historical recordings, their training had to be adapted to perform well

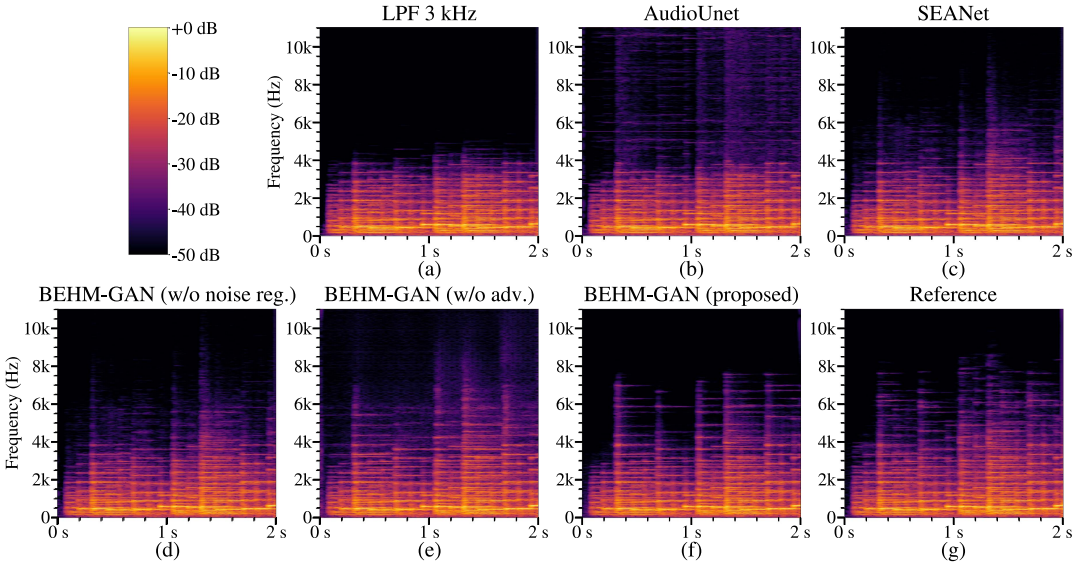

Fig. 7. Spectrograms of (a) a lowpass filtered reference signal, (b), (c), (d), (e), (f) its bandwidth-extended versions, and (g) a reference modern piano recording.

on this task. Unless otherwise specified, the compared baselines were trained using the same methodology as the BEHM-GAN (Section IV), using the same lowpass filters and noise regularization.

AudioUnet is a supervised model based on a time-domain U-Net [11], which was originally presented to enhance both speech and piano music signals sampled at 16 kHz. The model is optimized using a reconstruction L2 loss between the resulting output waveform and the original unfiltered audio. We use the PyTorch implementation of the model released by Sulun and Davies.[2]

SEANet is a GAN-based model that was used for speech enhancement [60] and speech bandwidth extension [14]. Its generator is a time-domain U-Net with dilated convolutions at the intermediate layers. An effort was made to replicate the implementation details from the larger non-real-time model evaluated in [14]. Given that the original SEANet utilizes very similar, if not the same, time-domain discriminators as in this paper, we opted to train SEANet using the same training objective as ours. This gave us better performance than the "feature" loss the authors originally applied and allowed us to directly evaluate the effects of using a spectrogram-based generator versus a time-domain one.

We also experimented with TFiLM [35], the GAN-based HiFi-GAN [13], and the diffusion probabilistic NU-Wave models [15]. However, we did not obtain positive results using these methods for bandwidth-extending historical recordings. We hypothesize that, in contrast to speech, the aforementioned models may not be well suited for processing music instead of speech or that further hyperparameter optimization is necessary to adapt the models to our training methodology. We decided not to include these baselines in the formal evaluation to avoid reporting misleading results and to not overload the number of listening conditions in the subjective evaluation.

So as to understand the effects of the main components of our approach, we also include four ablated versions of our model in the formal evaluation. Firstly, we study the importance of noise regularization by training a model without adding white noise (w/o noise reg.). Secondly, we switch the adversarial training objective for an L2 reconstruction loss in the complex-spectrogram domain (w/o adv.). The multi-resolution STFT loss L_{rec} is not used, being ineffective if used alone, since the phase information in the spectrogram is ignored. We also report, although only with objective metrics, the effect of adding noise to the input audio signal during inference, in the same way as we do during training (with noise inf.). Finally, we also included in the objective evaluation a model based on a complex-aware architecture [47], which consists of the same architecture from Fig. 4, but replacing each of the convolutional blocks with their complex-valued counterparts.

Figs. 7 and 8 show a visual comparison of the spectrogram representations of the compared models in a modern-lowpass filtered example and an old recording, respectively. Figs. 7(a) and 8(b) present, respectively, the bandlimited input signals for the examples, and the other spectrograms visualize how each method recreates the missing high-frequency content. The reference signal is added in Fig. 7(g) so that the results can be compared with the original real spectrogram. As is evident, some of the compared methods produce a more realistic spectra than others. A reference signal is unavailable for the old recording in Fig. 8. In this case, we additionally present the spectrogram of the noisy old recording before applying the denoiser in Fig. 8(a), which reveals that the old recordings contains only noise and distortion above about 4 kHz.

Figs. 7(f) and 8(g) show the spectrogram of the results of the proposed method. By comparing them with the input bandlimited signals (Figs. 7(a) and 8(b), respectively), one can observe

[2][Online]. Available: https://github.com/serkansulun/deep-music-enhancer

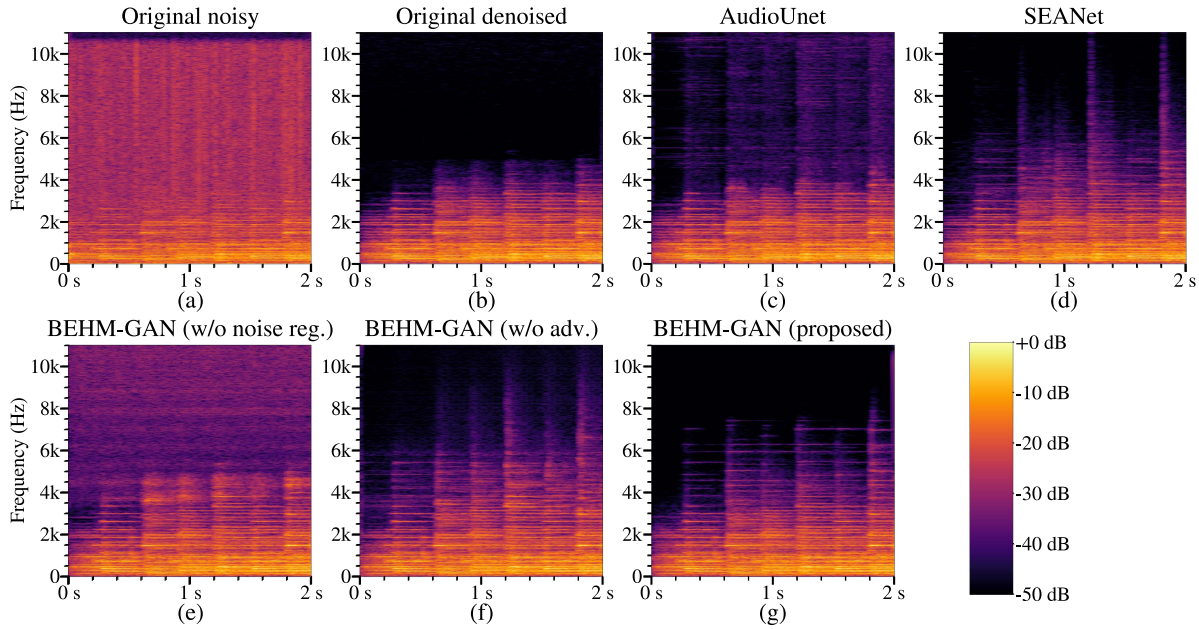

Fig. 8. Spectrograms of (a) an original noisy historical recording, (b) its a denoised version, and (c), (d), (e), (f), (g) the bandwidth-extended versions of (b).

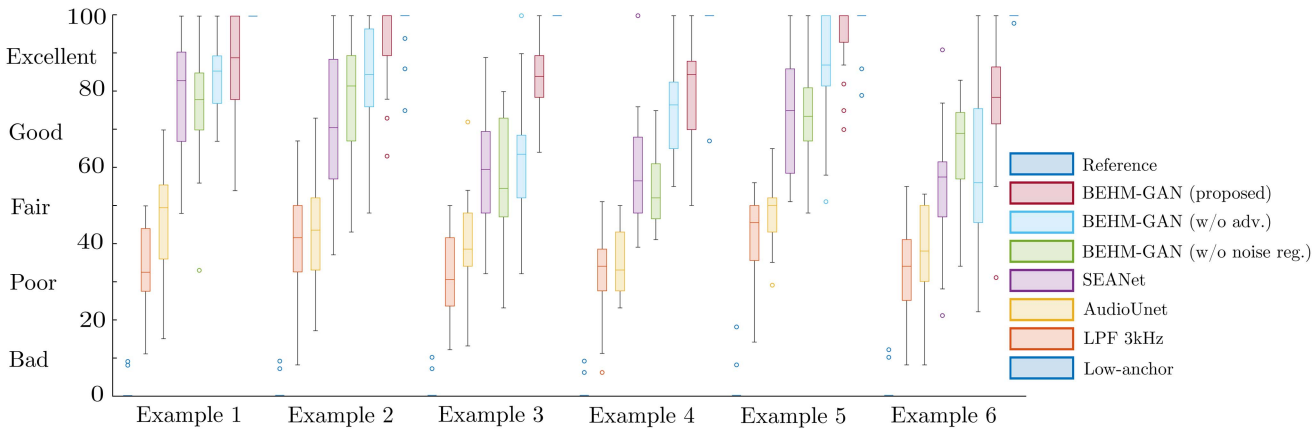

Fig. 9. Box-plot visualizations of the listening test results on synthetic lowpass filtered recordings.

how the bandwidth has been practically doubled. Nevertheless, almost no new frequency has been generated above 8 kHz. Another issue that can be perceived is the fact that the model is sometimes too aggressive in building up high frequencies, meaning that sometimes they may start too early. We hypothesize that the problem may come from the time-frequency processing, as the FFT window may be smearing the transients of the piano notes.

B. Objective Evaluation

To conduct the objective evaluation, we utilize the test set described in Section IV-C, which comprises 70 min. of modern broadband classical music recordings. All the audio signals from the test set were resampled at the rate $f_s = 22.05$ kHz and were split into non-overlapping 5-s segments. A sixth-order Butterworth lowpass filter is applied at the fixed cutoff frequency of 2 kHz, 3 kHz, and 4 kHz to imitate different bandwidth limitations. Note that the testing filters are purposely different from the training filters in order to evaluate the models in out-of-distribution filtering conditions. The magnitude responses of the three Butterworth filters used for testing are shown in Fig. 5, together with the filters used during training.

The proposed method and the aforementioned baselines are evaluated using three objective metrics: log-spectral distance (LSD), VGG distance (VGG), and Fréchet Audio Distance (FAD) [61].

1) Log-Spectral Distance: The LSD, a frequency-domain metric that has been popularly used in bandwidth-extension literature [11], [12], [13], [42], is defined as:

$$\text{LSD} = \frac{1}{T}\sum_{t=1}^{T}\sqrt{\frac{1}{K}\sum_{k=1}^{K}\left(\log|Y_{t,k}|^2 - \log|\hat{Y}_{t,k}|^2\right)^2}, \quad (7)$$

TABLE I
OBJECTIVE METRICS, WHERE LOWER IS BETTER FOR ALL CASES. THE BEST RESULT IN EACH COLUMN IS HIGLIGHTED

	$f_c = 2$ kHz			$f_c = 3$ kHz			$f_c = 4$ kHz			Historical recording
	LSD	**VGGish**	**FAD**	**LSD**	**VGGish**	**FAD**	**LSD**	**VGGish**	**FAD**	**FAD**
LPF/Original	1.01	5.38	4.96	0.82	3.93	3.48	0.71	3.43	2.07	2.16
AudioUnet	0.99	4.57	3.51	0.84	3.99	1.96	0.79	3.73	1.56	2.88
SEANet	0.99	4.41	3.87	0.81	3.38	1.18	0.75	3.21	1.05	1.66
BEHM-GAN (proposed)	0.87	4.15	3.44	**0.71**	3.21	0.70	**0.66**	3.01	0.70	1.26
w/o noise reg.	0.92	4.45	5.14	0.73	3.54	2.60	0.7	3.47	1.04	6.00
w/o adv.	1.00	5.42	5.18	0.86	3.99	2.95	0.75	4.02	1.91	2.23
with complex conv.	**0.85**	4.21	**1.02**	**0.71**	3.26	**0.53**	**0.66**	2.93	**0.47**	1.67
with noise inference	**0.85**	**3.98**	3.34	**0.71**	**3.00**	0.68	**0.66**	**2.76**	0.64	**1.12**
Reference	-	-	0.58	-	-	0.58	-	-	0.58	-

where $Y_{t,k} = \text{STFT}(y)$ and $\hat{Y}_{t,k} = \text{STFT}(\hat{y})$ are the STFTs of the reference y and the bandwidth-extended audio signal $\hat{y}$, respectively. For the STFT computation, an analysis window of $K = 2048$ samples and a hop length of 512 samples is used.

2) VGG Distance: This metric is defined as the L2 distance between pairs of individual embeddings given by the VGGish network [62]. The VGGish network has been pre-trained for large-scale audio classification. Thus, this metric is expected to provide a distance measure focusing on the higher level features of the audio data. This metric was previously used to evaluate music denoising [4] and bandwidth-extension [16] methods.

3) Fréchet Audio Distance: FAD [61] has been adapted for audio from the Fréchet Inception Distance (FID), a frequently used metric to evaluate image-generative models. This metric also uses the VGGish embeddings to compare the statistics between two collections of audio. FAD fits a multivariate normal distribution to a collection of background $\mathcal{N}(\mu_b, \Sigma_b)$ and evaluation $\mathcal{N}(\mu_e, \Sigma_e)$ embeddings. Then, the Fréchet distance between both distributions is defined as:

$$\text{FAD} = \|\mu_b - \mu_e\|_2 + \text{tr}(\Sigma_b + \Sigma_e - 2\sqrt{\Sigma_b \Sigma_e}), \quad (8)$$

where $\|\cdot\|_2$ is the L2 norm and $\text{tr}(\cdot)$ is the trace of a matrix. Being reference-free, we avoid computing FAD on paired data by dividing the test set into two equally-sized splits. One of them is used to compute the background statistics, whereas the second is used to evaluate the different models.

As FAD is reference-free, it is also used to evaluate the bandwidth-extension performance in real historical recordings. Similarly, the FAD is computed from 15 min of historical recordings using the same background statistics. The six historical recordings that we use for evaluation were extracted from "The Great 78 Project" [55], as mentioned in Section IV-C. However, since the FAD has the limitation of working at the sampling frequency of 16 kHz, which differs from that at which the BEHM-GAN operates (22.05 kHz), the audio signals must be resampled before computing this metric. As a consequence, the FAD only observes frequency components up to the Nyquist limit of 8 kHz, missing some high-frequency details. The study of the design of broadband reference-free audio quality metrics is left as future work.

No SNR-related metric is used for the objective evaluation because they are extremely sensitive to phase misalignments between pairs of data. Since the training and testing filters have a different phase response, the waveform representations of the reference and the bandwidth-extended audio can differ greatly, although they may sound similar. For this reason, the SNR results do not correlate with perceptual audio quality and are discarded.The objective results are tabulated in Table I. Note that a reference condition for the FAD metrics is added in the lowpass filtered results. This refers to computing the FAD between the two test splits of broadband piano recordings. Thus, the reference results can be considered as a lower bound of the FAD metric.

TABLE II
COMPARISON STUDY OF ADDING NOISE AT THE INFERENCE STAGE. THE BEST RESULT IN EACH ROW IS HIGHLIGHTED

		FAD	
		w/o noise at inf.	with noise at inf.
Lowpass filtered $f_c = 3$ kHz	AudioUnet	**1.96**	3.64
	SEANet	1.18	**1.09**
	BEHM-GAN	0.70	**0.68**
Historical recordings	AudioUnet	**2.88**	6.74
	SEANet	**1.66**	1.91
	BEHM-GAN	1.26	**1.12**

The proposed method outperforms the two compared baselines in all the evaluated conditions and, most importantly, improves over the "LPF/Original" condition, where no bandwidth extension was applied. The performance decreases when the noise regularization is not used, and also when the model is trained without the adversarial losses, proving that these are critical features of the model within the context of the results obtained. We will return to this point when analyzing the subjective evaluation results in Section V-D.

The BEHM-GAN obtains the most substantial improvement when $f_c = 3$ kHz, which corresponds to the mean cutoff frequency used for the training filters. The model also generalizes well when the cutoff frequency is higher ($f_c = 4$ kHz). However, when decreasing the cutoff frequency ($f_c = 2$ kHz), the metrics deteriorate a little as the task gets considerably harder. Nevertheless, even in this case, a significant improvement is seen in all three metrics with respect to the unprocessed lowpass filtered condition. The objective metrics also show that the FAD is decreased by a factor of two when the BEHM-GAN is evaluated with real historical recordings, implying that the model is able to generalize in this real-world case.

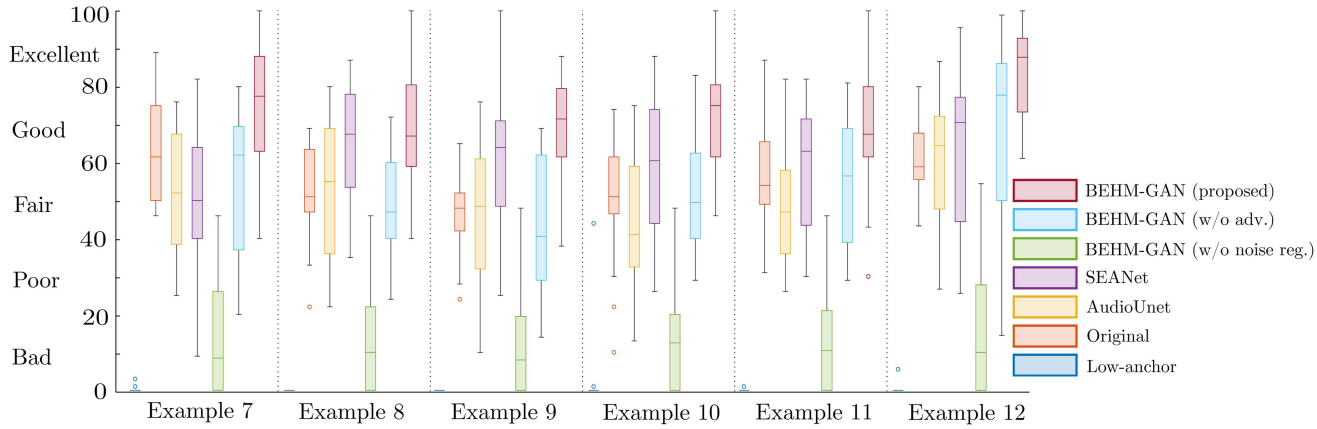

Fig. 10. Box-plot visualizations of the listening test results on real historical recordings.

Substituting the convolutional layers for complex convolutions [47] does not produce an improvement in the reference-based objective metrics (LSD and VGGish). This configuration improves the FAD score, but only for the synthetic filtered data and not for the real historical recordings. In informal listening, we did not find any perceptual improvement in the results. As a consequence, we opt to use the double-real representation instead, given that the complex-valued layers require twice the amount of computation.

The metrics of the proposed method consistently improve when noise regularization is added at the inference stage. This is an expected result given that, in this case, the test example gets closer to the training distribution. However, as seen in Table II, this does not necessarily happen with the other baseline methods. We noticed that AudioUnet and, to a lesser extent, SEANet sometimes do not succeed at completely suppressing the extra added Gaussian noise. No perceptual improvement was observed in the BEHM-GAN with or without noise added during inference or not. As a consequence, to make a fair comparison, only versions with no extra noise are included in the subjective evaluations.

C. *Subjective Evaluation of Synthetic Filtered Data*

Since finding an objective metric that correlates with human perception is not easy, a formal subjective evaluation is conducted. We designed a blind listening test structured as two consecutive sessions, one to assess the performance of our method with simulated lowpass filtered piano music and the other to evaluate the performance in real historical piano recordings, which is detailed in Section V-D.

The first part was designed following the MUSHRA recommendation [63] with the purpose to evaluate the bandwidth-extension performance in synthetic lowpass filtered recordings. The listeners had to grade, on a perceptual scale from 0 to 100, the audio quality in eight different conditions. The reference signal was a contemporary broadband recording, expected to be rated as 100. Another condition was a lowpass filtered version of the reference at the cutoff frequency of 3 kHz (LPF 3 kHz), applying the same Butterworth filter that was used in the objective evaluation. The rest of the conditions were five different bandwidth-extended versions of the lowpass filtered signal, using the same methods as in Section V-B. Also included as a low anchor is an easy-to-recognize poor-quality signal lowpass filtered at 1.5 kHz, which was expected to be graded as 0 by the listeners. We included six different 10-s piano music examples, repeated twice in random order, forming a total of 12 pages in the MUSHRA test. The audio examples included in the test are available listen at the companion webpage.[3]

Altogether, 13 listeners participated in the listening test. However, one participant was discarded from the first part because they did not identify the reference in more than 15% of the occasions, as recommended in [63]. All subjects had previous experience in formal listening tests, three of them were female, and their average age was 29 years. None of the participants reported of known hearing defects. The two test sessions took, on average, 45 min to complete. The experiment was conducted in the sound-proof listening booths of the Aalto Acoustics Lab, providing the same isolated listening conditions for all subjects. The listening test was implemented using the webMUSHRA interface [64] that allowed the listeners to set loops if they wanted to focus on particular short passages of the audio signal. This feature was particularly useful for some participants, since the most noticeable differences between the conditions were localized in certain details, such as at the attack transients of the piano tones.

The results of the first session are plotted in Fig. 9. Table III presents the distances between the median scores of the BEHM-GAN and the compared conditions. The level of statistical significance given by a paired t-test is also marked. Examples 3, 4, and 6 contained intense *fortissimo* piano passages, where the effect of the bandwidth limitation was easily audible. This explains why all the compared models obtained consistently lower ratings in these examples than in Examples 1, 2, and 5, which were softer and contained less high-frequency content.

For all six examples, the proposed method obtained higher median scores than the other evaluated conditions. As indicated

[3][Online]. Available: http://research.spa.aalto.fi/publications/papers/ieee-taslp-behm-gan/

TABLE III
DIFFERENCES BETWEEN THE MEDIAN SCORES OF THE LISTENING TEST ON SYNTHETIC FILTERED RECORDINGS

	Synthetic Lowpassed Recordings					
	Ex. 1	Ex. 2	Ex. 3	Ex. 4	Ex. 5	Ex. 6
Low-anchor	89 ***	100 ***	84 ***	84.5 ***	100 ***	78.5 ***
LPF 3kHz	56.5 ***	58.5 ***	53.5 ***	50.5 ***	54.5 ***	44.5 ***
AudioUnet	39.5 ***	56.5 ***	45.5 ***	51.5 ***	50 ***	40.5 ***
SEANet	6	29.5 ***	24.5 ***	28 ***	25 ***	21 ***
BEHM-GAN (w/o noise reg.)	11 *	18.5 **	29.5 ***	32.5 ***	26.5 ***	9.5 **
BEHM-GAN (w/o adv.)	3.5	15.5 *	20.5 ***	8	13 *	22.5 ***
Reference	-15 ***	0	-16 ***	-15.5 ***	0	-21.5 ***

TABLE IV
DIFFERENCES BETWEEN THE MEDIAN SCORES OF THE LISTENING TEST ON REAL HISTORICAL RECORDINGS

	Real Historical Recordings					
	Ex. 7	Ex. 8	Ex. 9	Ex. 10	Ex. 11	Ex. 12
Low-anchor	77.5 ***	67 ***	71.5 ***	75 ***	67.5 ***	79 ***
Original (denoised)	16 *	16 ***	23.5 ***	24 ***	13.5 *	26 ***
AudioUnet	25.5 ***	12 **	23 ***	34 ***	20.5 ***	21 ***
SEANet	27.5 ***	-0.5	7.5	14.5 **	4.5 *	15.5 ***
BEHM-GAN (w/o noise reg.)	69 ***	57 ***	63.5 ***	62.5 ***	57 ***	70 ***
BEHM-GAN (w/o adv.)	15.5 ***	20 ***	31 ***	25.5 ***	11 **	9 *

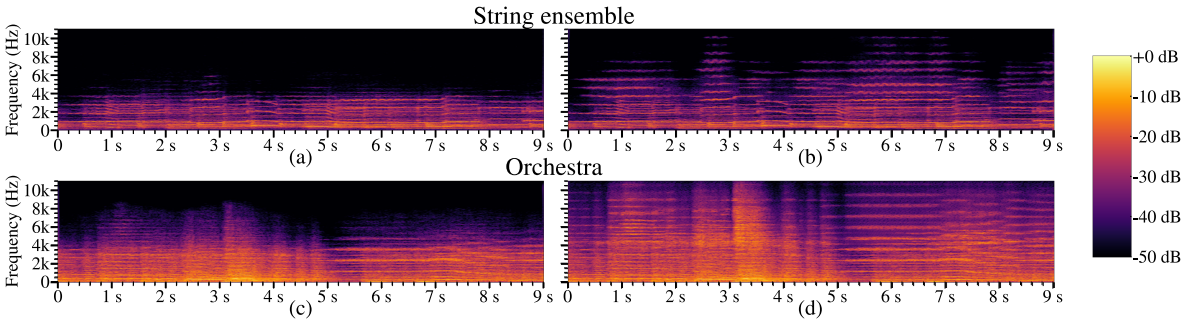

Fig. 11. Spectrograms of (a), (c) denoised historical recordings and (b), (d) their bandwidth-extended versions.

in paired t-tests (see Table III), most differences are statistically significant. The subjects easily identified the reference in all the examples except examples 2 and 5, where a large proportion of listeners rated the BEHM-GAN with the maximum score of 100.

The worst-rated model was AudioUnet, which introduced some annoying aliasing artifacts that can be seen at the upper part of Fig. 7(b). SEANet worked significantly better but still produced a slightly distorted sound. The ablated versions of the proposed method obtained relatively good scores but, in the majority of the cases, were outperformed by the full BEHM-GAN model.

D. Subjective Evaluation of Historical Recordings

The goal of the second session of the listening test was to evaluate the performance of the compared models in real historical recordings. The test method was a modified version of MUSHRA, where the audio examples were historical piano recordings. Since a broadband reference is unavailable, the reference presented to the listeners was, as in [65], the unprocessed signal, in our case a denoised bandlimited gramophone recording. The tested conditions were the various bandwidth-extended versions of the reference, using the same methods as in the previous session. The same low anchor was included with the expectation of it being graded as 0, forming a total of seven conditions. The participants were asked to grade the audio quality for each of the conditions on a scale from 0 to 100 with the same criteria as in the first session, where 100 corresponds to a hypothetical perfect version of the reference. Thus, the participants were discouraged to rate any of the conditions with a score of 100 unless the quality was considered enhanced in a perfectly realistic manner. The second session included six 10-s examples of historical piano recordings, also repeated twice, and the total number of pages was 12.

The experiment was conducted consecutively after the first session in the same conditions, with a short break between the two sessions. The test participants were also the same, except one who had to be discarded from the second part as they misunderstood the test question. This session took, on average, 20 min to complete.

The results of the second session are presented in Fig. 10 and Table IV. Given that a higher "excellent" anchor was not available, the resulting scores contain more variance. Despite

the wider confidence intervals, valuable conclusions can be extracted on how well each model generalizes to real historical recordings.

The BEHM-GAN obtained significantly better results than the original recording in all the examples, implying that the bandwidth extension improved the sound quality. The other compared models introduced some distortion artifacts that the listeners sometimes evaluated negatively. The proposed method obtained marginally better scores than the rest of the conditions in all the examples except example 8, where SEANet received a slightly higher median score. Nevertheless, the BEHM-GAN outperformed SEANet in four out of six examples, as is evident from Table IV. After finishing the test, some participants commented that the enhancement was often more noticeable with the time-domain SEANet model. However, the proposed method produced more realistic results, despite being more conservative.

Both ablated conditions suffered from a decline in performance. In particular, when noise regularization was ablated (BEHM-GAN w/o noise reg.), the resulting metrics were disappointing. These results can be explained by looking at the example in Fig. 8(e), which looks noisy and distorted. We hypothesize that the model learned to invert the frequency response of a hypothetical lowpass filter but it failed when it encountered an out-of-distribution example where no such filter was applied. By closely inspecting the spectrogram of Fig. 8(e), the model is observed to significantly boost the frequency bands above 4 kHz and introduce some horizontally-shaped noisy components that resemble the side lobes of the lowpass filters seen during training. These results show that noise regularization is critical for the good performance of our system. Another observation is that the non-adversarial condition (BEHM-GAN w/o adv.) obtained worse scores in this case, implying that the proposed adversarial training objective is highly beneficial to generate a more realistic enhancement in this out-of-distribution scenario.

VI. Conclusion

This paper proposes the BEHM-GAN, a method to extend the bandwidth of historical music. The proposed method is based on a generative adversarial network and combines a time-frequency-domain generator with multiple time-domain discriminators. The BEHM-GAN is trained in a self-supervised fashion using lowpass filters to simulate the bandwidth limitation of old recordings. With the intention of strengthening the robustness of our model, we regularize the training by randomizing the cutoff frequency of the filters and perturbing the filtered signal with a small amount of Gaussian white noise. The trained generator is designed to be incorporated as the second step in a music restoration pipeline, where the first step is a deep music denoiser [17]. This is, to the best of our knowledge, the first successful work that extends the bandwidth of historical music recordings.

The proposed method is evaluated using solo piano music, with objective and subjective metrics. As we show in App. B, the BEHM-GAN can also be applied to other types of music, such as orchestral music or string ensembles. However, this implies retraining the model with specialized data, as our attempts to train the BEHM-GAN with a broader range of music resulted in weaker performance. We leave as future work to study ways to allow the model to have better generalization capabilities without the need for retraining. Another limitation is that our method does not consider all the degradations present in old recordings. While the denoiser does a good job suppressing the additive disturbances and the BEHM-GAN reduces the bandwidth limitation, many other distortion artifacts remain untreated and present in the signal. Further work needs to be done to design a more robust method that addresses these issues.

Appendix A Median Scores of the Listening Tests

Tables III and IV show for each of the listening test sessions, the differences between the median scores of the proposed BEHM-GAN model and the rest of the conditions, respectively. Thus, larger positive values represent worst median scores, relative to the scores received by the proposed method. Asterisks $(*)$ denote significant differences in a paired t-test, where $*$, $**$ and $***$ respectively indicate p-values < 0.05, < 0.01 and < 0.001.

Appendix B Experiments With Different Musical Instruments

We have also experimented applying our model to other kinds of music having different musical instruments, more precisely string formations and orchestral music. The model has been retrained with a different dataset for each case. For the first case, 9.5 h of string ensemble recordings from the MusicNet dataset [54] have been used as training data and, for the orchestral music experiment, the training data comprised 7 h of freely-available modern orchestral recordings from The Internet Archive [56].

Fig. 11 shows the spectrogram representations of two original historical recordings and the results after applying the BEHM-GAN. Compared with solo piano recordings, strings and orchestral music recordings have a much richer high-frequency spectra. This implies that the bandwidth extension processing is often more noticeable in these cases. The BEHM-GAN seemed to perform well with string ensemble music. As can be observed in Fig. 11 b, the proposed method was able to extend the vibrato sound of a violin. In the case of orchestral music, the model succeeded in enhancing softer passages with strings and winds. However, in this case, we noticed some annoying artifacts with louder percussive instruments, such as drums and cymbals (Fig. 11 d). We attribute this weaker performance to the higher variance present in the training data, as orchestral music contains a wide ensemble of different instruments. This variance represents a higher difficulty for the model to generate more robust results. A set of audio examples is available for listening in the companion webpage.[4]

[4][Online]. Available: http://research.spa.aalto.fi/publications/papers/ieee-taslp-behm-gan/

ACKNOWLEDGMENT

The authors would like to thank the participants of the listening test. Special thanks go to Mr. Luis Costa for proofreading the manuscript. Additionally, the authors acknowledge the computational resources provided by the Aalto Science-IT project.

REFERENCES

[1] S. J. Godsill and P. J. W. Rayner, *Digital Audio Restoration–A Statistical Model Based Approach.* Berlin, Germany: Springer, 1998.

[2] P. A. A. Esquef, "Audio Restoration," in *Handbook of Signal Processing in Acoust*. New York, NY, USA: Springer, 2008, pp. 773–784.

[3] F. Rund, V. Vencovský, and M. Semanský, "An evaluation of click detection algorithms against the results of listening tests," *J. Audio Eng. Soc.*, vol. 69, no. 7/8, pp. 586–593, 2021.

[4] Y. Li, B. Gfeller, M. Tagliasacchi, and D. Roblek, "Learning to denoise historical music," in *Proc. 21st ISMIR Conf.*, 2020, pp. 504–511.

[5] N. Perraudin, N. Holighaus, P. Majdak, and P. Balazs, "Inpainting of long audio segments with similarity graphs," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 26, no. 6, pp. 1083–1094, Jun. 2018.

[6] A. Marafioti, P. Majdak, N. Holighaus, and N. Perraudin, "GACELA: A generative adversarial context encoder for long audio inpainting," *IEEE J. Sel. Topics Signal Process.*, vol. 15, no. 1, pp. 120–131, Jan. 2021.

[7] P. Záviška, P. Rajmic, A. Ozerov, and L. Rencker, "A survey and an extensive evaluation of popular audio declipping methods," *IEEE J. Sel. Topics Signal Process.*, vol. 15, no. 1, pp. 5–24, Jan. 2021.

[8] M. Miron and M. Davies, "High frequency magnitude spectrogram reconstruction for music mixtures using convolutional autoencoders," in *Proc. Int. Conf. Digit. Audio Effects*, 2018, pp. 173–180.

[9] M. Lagrange and F. Gontier, "Bandwidth extension of musical audio signals with no side information using dilated convolutional neural networks," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2020, pp. 801–805.

[10] S. Hu, B. Zhang, B. Liang, E. Zhao, and S. Lui, "Phase-aware music super-resolution using generative adversarial networks," in *Proc. Interspeech*, 2020, pp. 4074–4078.

[11] V. Kuleshov, S. Z. Enam, and S. Ermon, "Audio super resolution using neural nets," in *Proc. Int. Conf. Learn. Representations (Workshop Track)*, 2017.

[12] H. Wang and D. Wang, "Towards robust speech super-resolution," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 29, pp. 2058–2066, 2021.

[13] J. Su, Y. Wang, A. Finkelstein, and Z. Jin, "Bandwidth extension is all you need," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2021, pp. 696–700.

[14] Y. Li, M. Tagliasacchi, O. Rybakov, V. Ungureanu, and D. Roblek, "Real-time speech frequency bandwidth extension," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2021, pp. 691–695.

[15] J. Lee and S. Han, "NU-Wave: A diffusion probabilistic model for neural audio upsampling," in *Proc. Interspeech*, 2021, pp. 1634–1638.

[16] S. Sulun and M. E. Davies, "On filter generalization for music bandwidth extension using deep neural networks," *IEEE J. Sel. Topics Signal Process.*, vol. 15, no. 1, pp. 132–142, Jan. 2021.

[17] E. Moliner and V. Välimäki, "A two-stage U-Net for high-fidelity denoising of historical recordings," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2022, pp. 841–845.

[18] I. Goodfellow et al., "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.

[19] E. Larsen and R. Aarts, *Audio Bandwidth Extension: Application of Psychoacoustics, Signal Processing and Loudspeaker Design*. Hoboken, NJ, USA: Wiley, 2005.

[20] M. Nilsson and W. B. Kleijn, "Avoiding over-estimation in bandwidth extension of telephony speech," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2001, pp. 869–872.

[21] T. Ziegler, A. Ehret, P. Ekstrand, and M. Lutzky, "Enhancing MP3 with SBR: Features and capabilities of the new mp3PRO algorithm," in *Proc. Audio Eng. Soc. 112th Conv.*, 2002.

[22] Q. Huang, T. Liu, X. Wu, and T. Qu, "A generative adversarial net-based bandwidth extension method for audio compression," *J. Audio Eng. Soc.*, vol. 67, no. 12, pp. 986–993, 2019.

[23] J. Makhoul and M. Berouti, "High-frequency regeneration in speech coding systems," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 1979, pp. 428–431.

[24] J. Abel et al., "A subjective listening test of six different artificial bandwidth extension approaches in English, Chinese, German, and Korean," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2016, pp. 5915–5919.

[25] H. Carl and U. Heuter, "Bandwidth enhancement of narrow-band speech signals," in *Proc. Eur. Signal Process. Conf.*, 1994, pp. 1178–1181.

[26] M. Dietz, L. Liljeryd, K. Kjorling, and O. Kunz, "Spectral band replication, a novel approach in audio coding," in *Proc. Audio Eng. Soc. 112th Conv.*, 2002, pp. 49–65.

[27] K.-Y. Park and H. S. Kim, "Narrowband to wideband conversion of speech using GMM based transformation," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2000, pp. 1843–1846.

[28] H. Seo, H.-G. Kang, and F. Soong, "A maximum a posterior-based reconstruction approach to speech bandwidth expansion in noise," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2014, pp. 6087–6091.

[29] P. Jax and P. Vary, "Artificial bandwidth extension of speech signals using MMSE estimation based on a hidden Markov model," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2003, pp. 680–683.

[30] J. Kontio, L. Laaksonen, and P. Alku, "Neural network-based artificial bandwidth expansion of speech," *IEEE Trans. Audio Speech Lang. Process.*, vol. 15, no. 3, pp. 873–881, Mar. 2007.

[31] 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 Lang. Process.*, vol. 19, no. 7, pp. 2170–2183, Sep. 2011.

[32] K. Li and C.-H. Lee, "A deep neural network approach to speech bandwidth expansion," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2015, pp. 4395–4399.

[33] K. Li, Z. Huang, Y. Xu, and C. Lee, "DNN-based speech bandwidth expansion and its application to adding high-frequency missing features for automatic speech recognition of narrowband speech," in *Proc. Interspeech*, 2015, pp. 2578–2582.

[34] A. Gupta, B. Shillingford, Y. Assael, and T. C. Walters, "Speech bandwidth extension with wavenet," in *Proc. IEEE Work. Appl. Signal Process. Audio Acoust.*, 2019, pp. 205–208.

[35] S. Birnbaum, V. Kuleshov, Z. Enam, P. W. Koh, and S. Ermon, "Temporal FiLM: Capturing long-range sequence dependencies with feature-wise modulations," in *Proc. Adv. Neural Inf. Process. Syst.*, 2021.

[36] T. Y. Lim, R. A. Yeh, Y. Xu, M. N. Do, and M. Hasegawa-Johnson, "Time-frequency networks for audio super-resolution," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2018, pp. 646–650.

[37] J. Lin, Y. Wang, K. Kalgaonkar, G. Keren, D. Zhang, and C. Fuegen, "A two-stage approach to speech bandwidth extension," in *Proc. Interspeech*, 2021, pp. 1689–1693.

[38] K. Zhang, Y. Ren, C. Xu, and Z. Zhao, "WSRGlow: A glow-based waveform generative model for audio, super-resolution," in *Proc. Interspeech*, 2021, pp. 1649–1653.

[39] M. Mirza and S. Osindero, "Conditional generative adversarial nets," 2014, *arXiv:1411.1784*.

[40] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1125–1134.

[41] S. Kim and V. Sathe, "Bandwidth extension on raw audio via generative adversarial networks," 2019, *arXiv:1903.09027*.

[42] S. E. Eskimez, K. Koishida, and Z. Duan, "Adversarial training for speech super-resolution," *IEEE J. Sel. Topics Signal Process.*, vol. 13, no. 2, pp. 347–358, May 2019.

[43] R. Geirhos et al., "Shortcut learning in deep neural networks," *Nature Mach. Intell.*, vol. 2, no. 11, pp. 665–673, 2020.

[44] V.-A. Nguyen, A. H. Nguyen, and A. W. Khong, "TUNet: A. block-online bandwidth extension model based on transformers and self-supervised pretraining," in *Proc. IEEE Conf. Acoust. Speech Signal Process.*, 2022, pp. 161–165.

[45] P. Copeland, *Manual of Analogue Sound Restoration Techniques*. London, U.K.: British Library, 2008.

[46] C. Hummersone, "IoSR matlab toolbox," 2017. [Online]. Available: https://github.com/iosr-surrey/matlabtoolbox

[47] C. Trabelsi et al., "Deep complex networks," in *Proc. Int. Conf. Learn. Represent.*, 2018, pp. 6885–6889.

[48] D.-A. Clevert, T. Unterthiner, and S. Hochreiter, "Fast and accurate deep network learning by exponential linear units (ELUs)," in *Proc. Int. Conf. Learn. Represent.*, 2016. [Online]. Available: http://arxiv.org/abs/1511.07289

[49] U. Isik, R. Giri, N. Phansalkar, J.-M. Valin, K. Helwani, and A. Krishnaswamy, "PoCoNet: Better speech enhancement with frequency-positional embeddings, semi-supervised conversational data, and biased loss," in *Proc. Interspeech*, 2020, pp. 2487–2491.

[50] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2017, pp. 4700–4708.
[51] K. Kumar et al., "MelGAN: Generative adversarial networks for conditional waveform synthesis," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 14910–14921.
[52] X. Mao, Q. Li, H. Xie, R. Y. Lau, Z. Wang, and S. Paul Smolley, "Least squares generative adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2794–2802.
[53] R. Yamamoto, E. Song, and J.-M. Kim, "Parallel WaveGAN: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2020, pp. 6199–6203.
[54] J. Thickstun, Z. Harchaoui, and S. M. Kakade, "Learning features of music from scratch," in *Proc. Int. Conf. Learn. Represent.*, 2017.
[55] "The great 78 project.," Accessed: Feb. 11, 2023. [Online]. Available: https://great78.archive.org
[56] "The internet archive.," Accessed: Feb. 11, 2023. [Online]. Available: https://archive.org
[57] C. M. Bishop, "Training with noise is equivalent to Tikhonov regularization," *Neural Computation*, vol. 7, no. 1, pp. 108–116, Jan. 1995.
[58] T. Salimans and D. P. Kingma, "Weight normalization: A simple reparameterization to accelerate training of deep neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 901–909.
[59] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Represent.*, 2015.
[60] M. Tagliasacchi, Y. Li, K. Misiunas, and D. Roblek, "SEANet: A multi-modal speech enhancement network," in *Proc. Interspeech*, 2020, pp. 1126–1130.
[61] K. Kilgour, M. Zuluaga, D. Roblek, and M. Sharifi, "Fréchet audio distance: A reference-free metric for evaluating music enhancement algorithms.," in *Proc. Interspeech*, 2019, pp. 2350–2354.
[62] S. Hershey et al., "CNN architectures for large-scale audio classification," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2017, pp. 131–135.
[63] ITU-R, "Method for the subjective assessment of intermediate quality level of audio systems," *Int. Telecommun. Union*, Geneva, Switzerland, Tech. Rep. ITU-R BS.1534-3, Oct. 2015.
[64] M. Schoeffler et al., "webMUSHRA—A comprehensive framework for web-based listening tests," *J. Open Res. Softw.*, vol. 6, no. 1, pp. 1–8, 2018.
[65] E. Damskägg and V. Välimäki, "Audio time stretching using fuzzy classification of spectral bins," *Appl. Sci.*, vol. 7, no. 12, 2017, Art. no. 1293.

Eloi Moliner received the B.Sc. degree in telecommunications technologies and services engineering and the M.Sc. degree in telecommunications engineering from the Polytechnic University of Catalonia, Barcelona, Spain, in 2018 and 2021, respectively. He is currently a Doctoral Candidate with the Acoustics Lab, Aalto University, Espoo, Finland. His research interests include digital audio restoration and audio applications of machine learning.

Vesa Välimäki (Fellow, IEEE) received the M.Sc. and D.Sc. degrees in electrical engineering from the Helsinki University of Technology (TKK), Espoo, Finland, in 1992 and 1995, respectively. In 1996, he was a Postdoctoral Research Fellow with the University of Westminster, London, U.K. During 1997–2001, he was a Senior Assistant (cf. Assistant Professor) with TKK. During 2001–2002, he was a Professor of signal processing with the Pori unit of the Tampere University of Technology, Tampere, Finland. During 2008–2009, he was a Visiting Scholar with Stanford University, Stanford, CA, USA. He is currently a Full Professor of audio signal processing and the Vice Dean for Research in electrical engineering with Aalto University, Espoo. His research interests include audio and musical applications of signal processing and machine learning. Prof. Välimäki is a Fellow of the Audio Engineering Society. During 2007–2013, he was a Member of the Audio and Acoustic Signal Processing Technical Committee of the IEEE Signal Processing Society and is currently an Associate Member. During 2005–2009, he was an Associate Editor for the IEEE SIGNAL PROCESSING LETTERS, and during 2007–2011, he was and Associate Editor for the IEEE TRANSACTIONS ON AUDIO, SPEECH AND LANGUAGE PROCESSING. During 2015–2020, he was a Senior Area Editor of the IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH AND LANGUAGE PROCESSING. In 2007, 2015, and 2019, he was the Guest Editor of special issues of the *IEEE Signal Processing Magazine*, and in 2010, of a special issue of the IEEE TRANSACTIONS ON AUDIO, SPEECH AND LANGUAGE PROCESSING. He is currently the Editor-in-Chief of the *Journal of the Audio Engineering Society*.