EBEN: Extreme bandwidth extension network applied to speech signals captured with noise-resilient body-conduction microphones
Julien Hauret, Thomas Joubaud, Véronique Zimpfer, Éric Bavu

To cite this version:

HAL Id: hal-03830649
https://hal.science/hal-03830649
Submitted on 20 Mar 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
EBEN: EXTREME BANDWIDTH EXTENSION NETWORK APPLIED TO SPEECH SIGNALS CAPTURED WITH NOISE-RESILIENT BODY-CONDUCTION MICROPHONES

Julien Hauret† Thomas Joubaud† Véronique Zimpfer† Éric Bavu⋆

⋆ LMSSC, Conservatoire national des arts et métiers, Paris, France, HESAM Université
† Department of Acoustics and Soldier Protection, French-German Research Institute of Saint-Louis (ISL)

ABSTRACT
In this paper, we present Extreme Bandwidth Extension Network (EBEN), a Generative Adversarial network (GAN) that enhances audio measured with body-conduction microphones. This type of capture equipment suppresses ambient noise at the expense of speech bandwidth, thereby requiring signal enhancement techniques to recover the wideband speech signal. EBEN leverages a multiband decomposition of the raw captured speech to decrease the data time-domain dimensions, and give better control over the full-band signal. This multiband representation is fed to a U-Net-like model, which adopts a combination of feature and adversarial losses to recover an enhanced audio signal. We also benefit from this original representation in the proposed discriminator architecture. Our approach can achieve state-of-the-art results with a lightweight generator and real-time compatible operation.

Index Terms— Speech enhancement, PQMF-banks, Bandwidth extension, Frugal AI, Signal Processing

1. INTRODUCTION
Speech capture for radio communications is traditionally performed using microphones located near the speaker’s lips. However, this sound capture is sensitive to ambient noise, which reduces the intelligibility of communications in high noise levels such as in industry or on the battlefield. Under extreme conditions, even differential microphones are unable to get rid of high-level surrounding noise. Unconventional voice pickup systems such as bone conduction transducers, laryngophones or in-ear microphones integrated into hearing protections have great potential in many applications. These systems have higher impedance that is matched only by tissues and bones vibrations contrary to air molecules’, making noise pollution almost imperceptible within captured speech. Yet, further research is needed to optimize the effective bandwidth of the captured speech signal as mid and high frequencies are missing due to the intrinsic low-pass characteristic of the biological pathway.

Since the desirable system is a two-way communication device, this entails real-time execution constraints, i.e. a short processing time (≤ 20 ms) to be indistinguishable to the human ear. Moreover, edge computations are required to guarantee low latency, which in turn necessitates lightweight architectures. These considerations also match frugal AI requirements. Finally, the developed model should be robust to gender, accent, and speech loudness.

To perform this audio super-resolution task, we opt for deep learning techniques since classical signal processing source-filter approaches have limited capabilities when the information has completely disappeared along a given frequency interval. Indeed, neural networks’ ability to extract relevant features for the downstream task allows matching high and low frequency contents. Therefore, human processing should be minimal to let the network build its own representation, and could benefit from raw waveform inputs instead of spectrogram, mel-spectrogram or MFCC. This trend is endorsed by several works for various audio tasks. The use of raw audio can also be combined with multiband processing to speed up inference as in RAVE. To pursue the objective of fast inference, a fully convolutional architecture has been preferred in [10, 11], eventually U-Net-like for audio-to-audio tasks [12, 13]. In addition, a simple reconstruction loss may be insufficient for conditional generation, producing unrealistic samples. As shown in [14, 15, 16], adversarial networks [17] can significantly improve the naturalness of produced sound.

Based on the above observations, we developed EBEN, a new deep learning model, inspired by Seanet [18] to infer mid and high frequencies from speech containing only low frequencies. As in the original paper, we use a generator that maps the degraded speech signal to an enhanced version. The generator is optimized to produce samples which are close to the reference while maintaining a certain degree of naturalness at different time scales. A multiband decomposition using Pseudo-Quadrature Mirror Filters (PQMF) [19] is applied to reduce the temporal dimension of input features and to focus the signal’s discrimination solely on high frequency bands. We therefore share a common goal with [20], but differ in methodology and addressed degradations.

We used recordings[1] from a non conventional in-ear pro-

1available to listen at [https://jhauret.github.io/eben] alongside source code, hyperparameters and model performances
totype in order to study the signal degradation. Focusing on capture-induced degradation of a specific device does not penalize the generality of our approach. This family of sensors consistently degrades speech in the same way, acting as a low-pass filter. Variations mostly occur in terms of cut-off frequency, attenuation, and lack of coherence at specific frequencies. Thus, to address any other system, it would be enough to have a matching dataset. The few minutes of recordings at our disposal being insufficient, we trained our model and several baselines \cite{11,16,18,20} with the French version of the Librispeech dataset \cite{21} with simulated degradations, hoping to later release a publicly available dataset of speech capture with noise-resilient body-conduction microphones.

2. METHODS

2.1. Degradation estimation

The selected equipment is an early prototype of in-ear transducer \cite{23} developed by ISL and Cotral Lab. This device takes advantage of the speaker’s hearing protection by being placed inside a custom-molded earplug, which increases communication capabilities in challenging and noisy environments. The captured signal mainly contains speech with no environmental noise. However, the acoustic path between the mouth and the transducers acts as a low pass filter: practically no relevant speech signal is picked up above a threshold frequency. This phenomenon can be further influenced by the occlusion effect due to the fitting of the individual protectors and some complex interactions with tissues. To evaluate this transfer function, the experimental protocol consists in capturing speech simultaneously with the in-ear transducer and a regular microphone placed in front of the speaker’s mouth under noise-free conditions. Since speech signals are not stationary, several short-window estimates were respectively used to produce an estimated degradation – under the first approximation of a linear transfer function – shown in Fig. 1(a) and the coherence function on Fig. 1(b).

![Fig. 1. Degradation analysis of the in-ear transducer](image)

The joint analysis of plots in Fig. 1 shows that speech frequency content is only captured in a range below 2 kHz: the coherence function is close to zero above this frequency, and the gain of the transfer function is very low. This indicates that the in-ear microphone exhibits a very high attenuation at mid and high frequencies. Interestingly, at very low frequencies, the coherence function also indicates a lack of correlation between the two signals, since in-ear transducers also pick up physiological sounds produced by swallowing or blood flow. Finally, two anti-resonances are observed at 900 Hz and 1700 Hz, corresponding to vibration nodes of the occluded inner ear and propagation in bones and tissues.

2.2. Pseudo Quadrature Mirror Filter formalism

The Quadrature Mirror Filter (QMF) banks, introduced in \cite{23}, are a set of analysis filters \( \{ H_i \} \subset [1, M] \) used to decompose a signal into several non-overlapping channels of same bandwidth, and synthesis filters \( \{ G_i \} \subset [1, M] \) used to recompose the signal afterward. Those filters are obtained from the same lowpass prototype filter. A typical frequency response for a M-band PQMF bank is given in Fig. 2. Fig. 3 shows the entire pipeline. The reconstruction is exact if \( \{ H_i \} \subset [1, M] \) and \( \{ G_i \} \subset [1, M] \) have an infinite support. In practice, a convolution kernel of \( 8M \) is enough to produce a near perfect pseudo reconstruction \( \hat{x} \) with well-chosen filters \cite{19}.

![Fig. 2. Frequency response of a PQMF filter bank](image)

When used in a subband coding context, the PQMF analysis filter’s outputs are decimated (downsampled) by a factor \( M \). This can speed up inference because the different bands are modeled as conditionally independent. The PQMF analysis can also be seen as a downsampling operator, allowing to generate channels with considerably reduced redundancy, leading to a reduction in computational complexity. Along with EBEN source code, we provide a modern and efficient implementation of the PQMF analysis and synthesis with native Pytorch functions, using strided convolutions and strided transposed convolutions only.

\footnote{with an optimization process to minimize the reconstruction error}
2.3. Model architecture

Unlike the Seanet generator, where the very first layer is a plain convolution used to process the audio at the original 16kHz sampling rate, we propose for EBEN to encapsulate a U-Net-like generator between a PQMF analysis layer and a PQMF synthesis layer. This allows to reduce the memory footprint of the model by reducing the first embeddings samplerate by a factor of M. It also makes possible to keep only subbands having voice content to feed the U-Net first convolution. Moreover, the number of encoder/decoder blocks is reduced to meet the constraints of real-time applications.

EBEN’s discriminator also differs from the Seanet approach, as we directly exploit the PQMF subbands as multiple discriminator’s inputs, without recombining nor upsampling the reconstructed subband signals. As shown in Fig 4, we adopt a multiscale ensemble discriminator approach, inspired by the work of Kumar et al. in [14], whose inputs are the upper bands of the PQMF decomposition. This allows the ensemble of discriminators to analyze the generated subband signals at different time scales, and helps the generator to be trained to generate high quality samples despite the fact that each discriminator is relatively simple. The subband discriminators \(D_k\), \(k \in \{1,2,3\}\) exhibit similar receptive fields to the original Seanet discriminator. As shown in Fig. 4, the full scale MelGAN discriminator was however kept from the Seanet approach, in order to ensure coherence between bands.

2.4. Loss functions

Discriminators are trained with a classical hinge loss. In \(\mathcal{L}_D\), and \(\mathcal{L}_G\), G represents the generator and \(D^{(l)}_k\) represents the layer \(l\) of the discriminator (among \(L_k\) layers) of scale \(k\) (among \(K\) scales) at time \(t\). \(F_{k,l}\) and \(T_{k,l}\) are the number of features and temporal length for given indices. \(x\) is the in-ear signal while \(y\) is the reference.

\[
\mathcal{L}_D = \mathbb{E}_y \left[ \frac{1}{K} \sum_{k \in [0,3]} \frac{1}{T_{k,L_k}} \sum_t \max(0,1-D_{k,t}(y)) \right] + \\
\mathbb{E}_x \left[ \frac{1}{K} \sum_{k \in [0,3]} \frac{1}{T_{k,L_k}} \sum_t \max(0,1+D_{k,t}(G(x))) \right]
\]

For the generator, we adopted a loss function composed of a generative part \(\mathcal{L}_G^{adv}\) and a reconstructive part \(\mathcal{L}_G^{rec}\), balanced by \(\lambda : \mathcal{L}_G = \mathcal{L}_G^{adv} + \lambda \mathcal{L}_G^{rec}\).

\[
\mathcal{L}_G^{adv} = \mathbb{E}_x \left[ \frac{1}{K} \sum_{k \in [0,3]} \frac{1}{T_{k,L_k}} \sum_t \max(0,1-D_{k,t}(G(x))) \right]
\]

\[
\mathcal{L}_G^{rec} = \mathbb{E}_x \left[ \frac{1}{K} \sum_{k \in [0,3]} \frac{1}{T_{k,L_k}} \sum_t \left\| D^{(l)}_{k,t}(y) - D^{(l)}_{k,t}(G(x)) \right\|_{L_1} \right]
\]

This combination allows to generate audio samples as close as possible to the reference signal thanks to \(\mathcal{L}_G^{rec}\), while remaining creative at high frequencies when no information is available in the degraded signal (especially for fricatives) thanks to \(\mathcal{L}_G^{adv}\). Note that \(\mathcal{L}_G^{adv}\) does not operate directly in the time domain but in discriminators domain to focus on the signal semantic (feature matching).

3. EXPERIMENTS

A simulated degradation (described in Sec. [3.1]) is performed on the French Librispeech [21] dataset. Different models [11, 16, 18, 20] and the EBEN approach have been trained on this imitated in-ear dataset during 2 days on a single RTX 2080 Ti GPU. Each model is evaluated objectively (Sec. [3.2]) and subjectively (Sec. [3.3]). No parameter tuning nor early stopping was performed.

3.1. Simulation of the dataset

The simulated degradation is based on an autoregressive moving-average model, using a 2nd order low-pass filter with a cutoff frequency of 600 Hz and unitary Q-factor that is applied using a \texttt{filtfilt} procedure to ensure zero phase shift. The superposition of the simulated and estimated degradation filters are represented on Fig. 1 (a). A Gaussian white noise with a power 23 dB below the selected signal is also added. This simplistic degradation still ensures a wide application field for developed algorithms and the ability to focus on the bandwidth extension issue. Future works will introduce variable filters with anti-resonance frequencies and realistic physiological noise.
3.2. Objective evaluation

To evaluate the model performances, Tab. 1 highlights several objective metrics: Perceptual Evaluation of Speech Quality (PESQ) [24], Scale-Invariant Signal-to-Distortion Ratio (SI-SDR) [25] and Short-Time Objective Intelligibility (STOI) [26] on test set. Speech enhancement being a one-to-many problem, these results should be analyzed cautiously. Indeed, a plausible signal with perfect intelligibility but still different from reference would be misjudged by the metrics. Note that these metrics are intrusive, since they require a groundtruth audio. Generally speaking, speech quality assessment is lacking non subjective and non comparative evaluation metrics. Works like Noresqua [27] attempted to introduce such non intrusive metrics, but were inefficient for our specific degradation. We also report on Tab. 1 the number of parameters for each model, showing that EBEN is among the lightest. Time latency and computational load are presented on the project website [2].

3.3. Subjective evaluation

To visually assess and compare the results of approaches, Fig. 5 shows some spectrograms obtained from the testing set. It can be observed that a purely reconstructive approach [11] is not sufficient to produce high frequencies. Indeed, when low frequency information is insufficient, the model predicts the mean of speech signals which is zero. Among generative approaches, our method reconstructs a fair amount of formants and minimizes artifacts. We also conducted a subjective evaluation of the different methods using the MUltiple Stimuli with Hidden Reference and Anchor [28] (MUSHRA) methodology on the GoListen platform [29]. Obtained results are also given on Tab. 1. MUSHRA-U is about ease of understanding while the MUSHRA-Q is about audio quality. Our approach is ranked first ex-aequo with Seanet for both aspects.

4. DISCUSSION AND FUTURE WORK

EBEN is a realtime-compatible network architecture to address the problem of bandwidth extension of speech signals captured with body-conduction microphones. The proposed multiband approach offers several benefits, including reduced parameters and the ability to select the bands to operate on with a GAN approach. Developing a lightweight causal architecture able to capture long-range dependencies to disambiguate some phonetical content could help to further improve speech reconstructions for severe degradations.

Acknowledgements: This work has been partially funded by the French National Research Agency under the ANR Grant No. ANR-20-THIA-0002. This work was also granted access to the HPC/AI resources of [CINES / IDRIS / TGCC] under the allocation 2022-AD011013469 made by GENCI.
5. REFERENCES


