# EBEN: EXTREME BANDWIDTH EXTENSION NETWORK APPLIED TO SPEECH SIGNALS CAPTURED WITH NOISE-RESILIENT 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 noise-resilient 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 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 [1, 2]. 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 ($\leq$ 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 [3] 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[4] or MFFC[5]. This trend is endorsed by several works [6, 7, 8] for various audio tasks. The use of raw audio can also be combined with multiband processing to speed up inference as in *RAVE* [9]. 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 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-

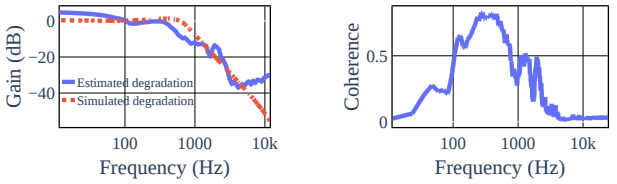
[1] available to listen at `https://jhauret.github.io/eben` alongside our source code and hyparameters

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 with the French version of the Librispeech dataset [21] with simulated degradations, hoping to later release a publicly available dataset of speech capture with non-conventional noise-resilient microphones.

## 2. METHODS

### 2.1. Degradation estimation

The selected equipment is a in-ear transducer prototype [22] 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. At first approximation, the propagation path is modeled as a linear transfer function. 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 shown in Fig. 1 (a) and the coherence function on Fig. 1(b).




(a) Transfer function (b) Coherence function

**Fig. 1**. Degradation analysis of the in-ear transducer

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 medium and high frequencies, and that the captured audio signal is essentially a mixture of analog and digital noise in this area. 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.

### 2.2. Pseudo Quadrature Mirror Filter formalism

The Quadrature Mirror Filter (QMF) banks, introduced in [23, 19], are a set of analysis filters $\{H_i\}_{i\in[1,M]}$ used to decompose a signal into several non-overlapping channels of same bandwidth, and synthesis filters $\{G_i\}_{i\in[1,M]}$ used to recompose the signal afterwards. 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, and Fig. 3 shows the entire pipeline. The reconstruction is exact if $\{H_i\}_{i\in[1,M]}$ and $\{G_i\}_{i\in[1,M]}$ 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*[2] filters.

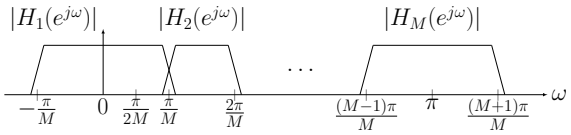



**Fig. 2**. Frequency response of the filter bank

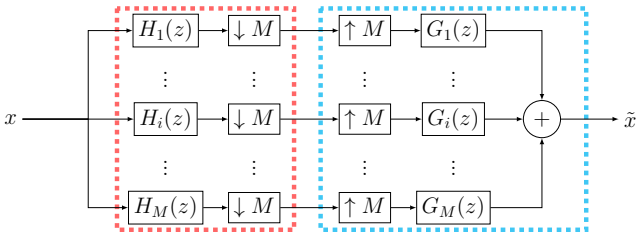



**Fig. 3**. PQMF Analysis and Synthesis : block-diagram

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 an 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.

[2] with an optimization process to minimize the reconstruction error

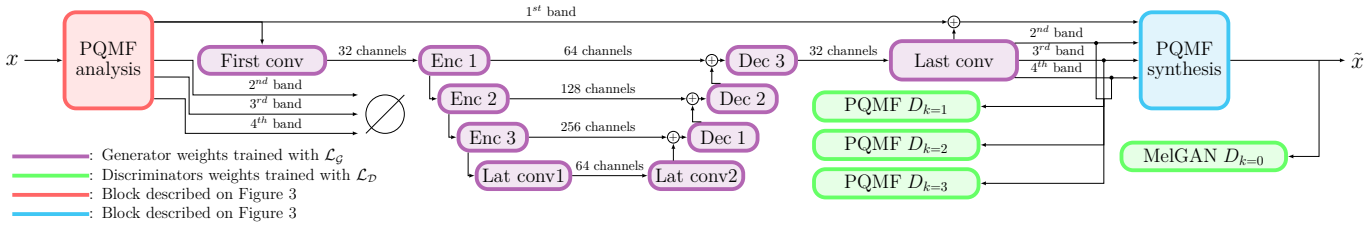



**Fig. 4**. Overall architecture represented for $M = 4$, one band with voice content fed to generator, others to PQMF discriminators

### 2.3. Model architecture

Unlike 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 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 input 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 $\{D_i\}_{i\in[1,2,3]}$ exhibit similar receptive fields to the original Seanet. 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 $\mathcal{L}_\mathcal{D}$, and $\mathcal{L}_\mathcal{G}$, $G$ represents the generator and $D^{(l)}_{k,t}$ 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.

$$\mathcal{L}_\mathcal{D} = E_y\left[\frac{1}{K}\sum_{k\in[0,3]}\frac{1}{T_{k,L_k}}\sum_t max(0, 1 - D_{k,t}(y))\right] + E_x\left[\frac{1}{K}\sum_{k\in[0,3]}\frac{1}{T_{k,L_k}}\sum_t max(0, 1 + D_{k,t}(G(x)))\right]$$

For the generator, we adopted a loss function composed of a generative part ${\mathcal{L}_\mathcal{G}}^{adv}$ (Eq. 2.4) and a reconstructive part ${\mathcal{L}_\mathcal{G}}^{rec}$ (Eq. 2.4), balanced by $\lambda$ : $\mathcal{L}_\mathcal{G} = {\mathcal{L}_\mathcal{G}}^{rec} + \lambda{\mathcal{L}_\mathcal{G}}^{adv}$.

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

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

This combination allows to generate audio samples as close as possible to the reference signal thanks to $\mathcal{L}^{rec}_\mathcal{G}$, while remaining creative at high frequencies when no information is available in the degraded signal (especially for fricatives) thanks to $\mathcal{L}^{adv}_\mathcal{G}$. Note that $\mathcal{L}^{rec}_\mathcal{G}$ 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 qualitatively (Sec. 3.2) and quantitatively (Sec. 3.3). No parameter tuning or early stopping were performed.

### 3.1. Simulation of the dataset

The simulated degradation is based on an autoregressive moving-average model, using a 2$^{\text{nd}}$ order low-pass filter with a cutoff frequency of 600 Hz and unitary Q-factor that is applied using a *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 corresponding to $0.5\%$ of 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.

| Speech \ Metrics | PESQ | SI-SDR | STOI | MUSHRA-i | MUSHRA-q | Gen params | Dis params |
|---|---|---|---|---|---|---|---|
| Simulated In-ear | 2.42 (0.34) | 8.4 (3.7) | 0.83 (0.05) | 50 (44) | 21 (18) | ∅ | ∅ |
| Audio U-net [11] | **2.24 (0.49)** | **11.9 (3.7)** | 0.87 (0.04) | 60 (37) | 32 (21) | 71.0 M | ∅ |
| Hifi-GAN v3[16] | 1.32 (0.16) | -25.1 (11.4) | 0.78 (0.04) | 40 (46) | 33 (30) | 1.5 M | 70.7 M |
| Seanet [18] | 1.92 (0.48) | 11.1 (3.0) | **0.89 (0.04)** | **80 (37)** | **80 (25)** | 8.3 M | 56.6 M |
| Strm-Seanet [20] | 2.01 (0.46) | 11.2 (3.6) | **0.89 (0.04)** | 68 (36) | 63 (28) | **0.7 M** | 56.6 M |
| EBEN (ours) | 2.08 (0.45) | 10.9 (3.3) | **0.89 (0.04)** | **78 (30)** | **79 (26)** | 1.9 M | **26.5 M** |

**Table 1**. Comparing models with: PESQ/SI-SDR/STOI on test set — MUSHRA-i (62 participants) and MUSHRA-q (60 participants) scores — number of parameters. Format is median (IQR). Significantly best values (acceptance=0.05) are in **bold**.

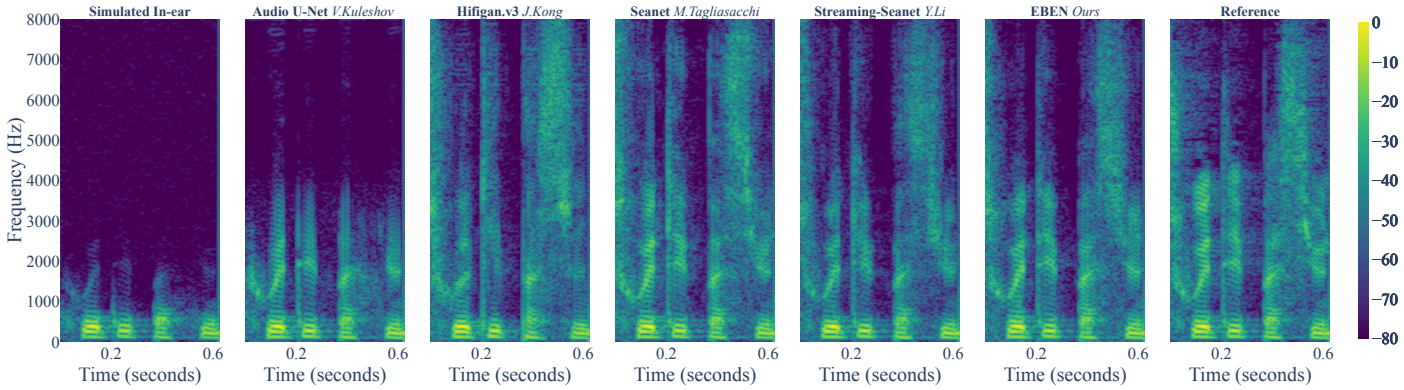



**Fig. 5**. Spectrograms of various bandwidth extension models sandwiched by the simulated in-ear and the reference signals.

### 3.2. Quantitative 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 both validation and test sets. 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 *Noresqa* [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.

### 3.3. Qualitative evaluation

To visually assess and compare the results of approaches, Fig. 5 shows some spectrograms. 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-i is about intelligibility while the MUSHRA-q is about audio quality. Our approach proves to be very competitive in both aspects.

## 4. DISCUSSION AND FUTURE WORK

This article describes EBEN, a realtime-compatible network architecture to address the problem of extreme bandwidth extension of speech signals captured with noise-resilient microphones. The proposed multiband approach offered several benefits, including reduced parameters and the ability to select the bands to operate on with a GAN approach. Still, the studied degradation is severe. Developping a lightweight causal architecture able to capture long-range dependencies to disambiguate some phonetical content could be part of the solution.

**Acknowledgements**: This work was granted access to the HPC/AI resources of [CINES / IDRIS / TGCC] under the allocation 2022-AD011013469 made by GENCI.

## 5. REFERENCES


[1] Jeffrey C Bos, David W Tack, and LCol Linda Bossi, "Speech input hardware investigation for future dismounted soldier computer systems," *DRCD Toronto CR*, vol. 64, pp. 2005, 2005.

[2] Barbara Acker-Mills, Adrianus Houtsma, and William Ahroon, "Speech intelligibility with acoustic and contact microphones," Tech. Rep., ARMY AEROMEDICAL RESEARCH LAB FORT RUCKER AL, 2005.

[3] Bernd Iser, Wolfgang Minker, and Gerhard Schmidt, *Bandwidth extension of speech signals*, Springer, 2008.

[4] Steven Davis and Paul Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE transactions on acoustics, speech, and signal processing*, vol. 28, no. 4, pp. 357–366, 1980.

[5] Bruce P Bogert, "The quefrency alanysis of time series for echoes; cepstrum, pseudo-autocovariance, cross-cepstrum and saphe cracking," *Time series analysis*, pp. 209–243, 1963.

[6] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in Neural Information Processing Systems*, vol. 33, pp. 12449–12460, 2020.

[7] Francois G Germain, Qifeng Chen, and Vladlen Koltun, "Speech denoising with deep feature losses," *arXiv preprint arXiv:1806.10522*, 2018.

[8] Wei Dai, Chia Dai, Shuhui Qu, Juncheng Li, and Samarjit Das, "Very deep convolutional neural networks for raw waveforms," in *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2017, pp. 421–425.

[9] Antoine Caillon and Philippe Esling, "Rave: A variational autoencoder for fast and high-quality neural audio synthesis," *arXiv preprint arXiv:2111.05011*, 2021.

[10] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu, "Wavenet: A generative model for raw audio," *arXiv preprint arXiv:1609.03499*, 2016.

[11] Volodymyr Kuleshov, S Zayd Enam, and Stefano Ermon, "Audio super resolution using neural networks," *arXiv preprint arXiv:1708.00853*, 2017.

[12] Daniel Stoller, Sebastian Ewert, and Simon Dixon, "Wave-u-net: A multi-scale neural network for end-to-end audio source separation," *arXiv preprint arXiv:1806.03185*, 2018.

[13] Amelie Bosca, Alexandre Guerin, Laureline Perotin, and Srdjan Kitic, "Dilated u-net based approach for multichannel speech enhancement from first-order ambisonics recordings," in *2020 28th European Signal Processing Conference (EUSIPCO)*. IEEE, 2021, pp. 216–220.

[14] Kundan Kumar, Rithesh Kumar, Thibault de Boissiere, Lucas Gestin, Wei Zhen Teoh, Jose Sotelo, Alexandre de Brébisson, Yoshua Bengio, and Aaron C Courville, "Melgan: Generative adversarial networks for conditional waveform synthesis," *Advances in neural information processing systems*, vol. 32, 2019.

[15] Sung Kim and Visvesh Sathe, "Bandwidth extension on raw audio via generative adversarial networks," *arXiv preprint arXiv:1903.09027*, 2019.

[16] Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae, "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," *Advances in Neural Information Processing Systems*, vol. 33, pp. 17022–17033, 2020.

[17] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, "Generative adversarial networks," *Communications of the ACM*, vol. 63, no. 11, pp. 139–144, 2020.

[18] Marco Tagliasacchi, Yunpeng Li, Karolis Misiunas, and Dominik Roblek, "Seanet: A multi-modal speech enhancement network," *arXiv preprint arXiv:2009.02095*, 2020.

[19] Truong Q Nguyen, "Near-perfect-reconstruction pseudo-qmf banks," *IEEE Transactions on signal processing*, vol. 42, no. 1, pp. 65–76, 1994.

[20] Yunpeng Li, Marco Tagliasacchi, Oleg Rybakov, Victor Ungureanu, and Dominik Roblek, "Real-time speech frequency bandwidth extension," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 691–695.

[21] Vineel Pratap, Qiantong Xu, Anuroop Sriram, Gabriel Synnaeve, and Ronan Collobert, "Mls: A large-scale multilingual dataset for speech research," *arXiv preprint arXiv:2012.03411*, 2020.

[22] "BIONEAR outstanding hearing for 4.0 professionals," `https://www.cotral-communication.com/en/industry.html`, Accessed: 2022-10-10.

[23] Joseph Rothweiler, "Polyphase quadrature filters–a new subband coding technique," in *ICASSP'83. IEEE International Conference on Acoustics, Speech, and Signal Processing*. IEEE, 1983, vol. 8, pp. 1280–1283.

[24] Antony W Rix, John G Beerends, Michael P Hollier, and Andries P Hekstra, "Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs," in *2001 IEEE international conference on acoustics, speech, and signal processing. Proceedings (Cat. No. 01CH37221)*. IEEE, 2001, vol. 2, pp. 749–752.

[25] Jonathan Le Roux, Scott Wisdom, Hakan Erdogan, and John R Hershey, "Sdr–half-baked or well done?," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 626–630.

[26] Cees H Taal, Richard C Hendriks, Richard Heusdens, and Jesper Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *2010 IEEE international conference on acoustics, speech and signal processing*. IEEE, 2010, pp. 4214–4217.

[27] Pranay Manocha, Buye Xu, and Anurag Kumar, "Noresqa: A framework for speech quality assessment using non-matching references," *Advances in Neural Information Processing Systems*, vol. 34, 2021.

[28] B Series, "Method for the subjective assessment of intermediate quality level of audio systems," *International Telecommunication Union Radiocommunication Assembly*, 2014.

[29] Dan Barry, Qijian Zhang, Pheobe Wenyi Sun, and Andrew Hines, "Go listen: an end-to-end online listening test platform," *Journal of Open Research Software*, vol. 9, no. 1, 2021.