Bandwidth extension of musical audio signals with no side information using dilated convolutional neural networks

Mathieu Lagrange, Félix Gontier

BANDWIDTH EXTENSION OF MUSICAL AUDIO SIGNALS WITH NO SIDE INFORMATION USING DILATED CONVOLUTIONAL NEURAL NETWORKS

Mathieu Lagrange, Félix Gontier

LS2N, CNRS, Centrale Nantes

ABSTRACT

Bandwidth extension has a long history in audio processing. While speech processing tools do not rely on side information, production-ready bandwidth extension tools of general audio signals rely on side information that has to be transmitted alongside the bitstream of the low frequency part, mostly because polyphonic music has a more complex and less predictable spectral structure than speech.

This paper studies the benefit of considering a dilated fully convolutional neural network to perform the bandwidth extension of musical audio signals with no side information on the magnitude spectra. Experimental evaluation using two public datasets, *medley-solos-db* and *gtzan*, respectively of monophonic and polyphonic music demonstrate that the proposed architecture achieves state of the art performance.

Index Terms— Artificial audio bandwidth extension, deep neural network, musical audio processing

1. INTRODUCTION

Bandwidth extension has a long standing history in telecommunication where the bitrate allowed by the given application may be reduced [1]. In that case, it is often beneficial to preserve a good perceptual quality in the lower frequencies, for example to preserve intelligibility in speech applications. It is thus useful to consider a processing unit on the receiver that is able to produce a wide-band signal in order to improve perceived quality given the narrow band signal consisting of the lower frequencies, typically up to 4 kHz for speech.

Many techniques have been introduced for speech, and most of them operate on the magnitude spectra, where the spectral envelope of the narrow band signal is used to predict the spectral envelope of the higher frequencies. Recent approaches consider deep neural networks to do so [2]. This approach assumes a source-filter model for speech production and requires some integration of two processing units, one responsible for the narrow band signal decoding and one responsible for the bandwidth extension. Extension on the magnitude spectra using a deep neural architecture has been proposed in [3]. Recurrent neural networks [4] or Wavenet architectures can also be considered [5]. In this case, the network directly predicts the wide band speech signal, given some information provided by the conditioning stack that processes the narrow band signal. This approach is very flexible, but still computationally demanding.

Due to interoperability requirements in telephony, the bandwidth extension process in speech is done without any side information. That is, the processing unit on the decoder side has to predict the higher frequency signal given some static knowledge and the lower frequency signal only.

In general audio coding, bandwidth extension has been introduced in the beginning of the millennium [6]. General audio coding is more complex than speech coding due to the variety of physical sources that may produce the signal that has to be encoded. Due to this, and the ability to control the whole transmission stack for most use cases, some side information is considered to perform the bandwidth extension process. This side information is computed using the wide band signal at the encoder side and transmitted within the bitstream. In [6], the main concept introduced is called spectral band replication (SBR) where the lower frequencies of the magnitude spectra are duplicated and transposed. Due to the typical exponential decay of magnitude with respect to the frequency, the overall magnitude of the transposed spectra has to be adjusted.

Some post processing steps can be undertaken to further improve the perceptual quality. As the encoder has access to the SBR prediction and the reference high frequency spectra, it is able to adapt to some special cases where considering the high frequency spectrum as the low frequency one will fail. For example, some high frequency tones that are perceptually salient may not be be recreated using the replication process. In this case, an additional processing unit can be considered to produce salient sinusoidal components [7]. The lower frequencies may have strong harmonics and the higher ones only noise like components. In this case, an inverse filtering is applied [8]. Extension for low delay applications have been proposed [9], as well as the application of the phase vocoder [10] to reduce unpleasant roughness typically introduced when considering SBR tools [11].

Learning approaches, such as non-negative matrix factorization [12], and deep learning approaches [13] have several benefits compared to algorithmic approaches discussed above. First, there may be no need for side information if the capacity of the model is sufficient to encode the rich relation-

Work partially funded by ANR CENSE

ship between the lower frequencies and the higher frequencies of the spectrum and if those encoded relationships are generic enough to produce satisfying results for real use case scenarios. Secondly, the relationships encoded by the model being non explicit, there is less chance of reaching a "glass-ceiling" in terms of perceptual quality.

To investigate further in this direction, we consider in this paper a deep convolutional network that operates in the magnitude spectrum. The model is described in Section 2. Its performance is evaluated using an experimental protocol described in Section 3 on two public datasets: the *medley-solos-db* and the *gtzan* datasets[1]. Outcomes of the performance analysis are described in Section 4 and discussed in Section 5.

Our main findings are that: 1) use of dilated convolutional filters lead to architectures that are less sensitive to the tuning of the other meta-parameter and reduce the complexity of the model while preserving a receptive field adapted to the task at hand, 2) compressive architectures like autoencoders do not perform favorably compared to a fully convolutional neural network without compression.

2. MODEL

The aim of a bandwidth extension system is to predict the high frequency part of the spectrum. In this paper, we consider audio data represented as magnitude spectra. The input and the output of the model consist in 128x10 (frequency x time) matrices that respectively represent the low frequency and high frequency parts of the audio for an approximate duration of 160ms (more details are provided in Section 3).

The architecture is a fully convolutional neural network [14] with L layers followed by rectified linear units (ReLU) activations, as described in Figure 1. The number of output convolution channels C is the same for all hidden layers. Convolution kernels also share the same size (K_t, K_f) in the time and frequency dimensions respectively. To keep the shape of magnitude spectra constant throughout the network, representations at the input of each layer are padded by replicating their boundary values depending on the kernel size. Dilated convolutions are considered to artificially increase the receptive field of the model [15], allowing it to capture patterns on larger scales without added parameters. From a signal processing standpoint, this procedure is equivalent to applying the convolution on a down-sampled version of the input. A fixed dilation ratio D is used in the frequency dimension for hidden layers, and no dilation is used in the input and output layers of the network. As a result each hidden layer increases the receptive field by $D(K-1)$ frequency bins compared to $K-1$ without dilation.

[1]The code is available at https://github.com/mathieulagrange/paperBandwidthExtensionCnn and some audio examples can be listened to at https://mathieulagrange.github.io/paperBandwidthExtensionCnn/demo.

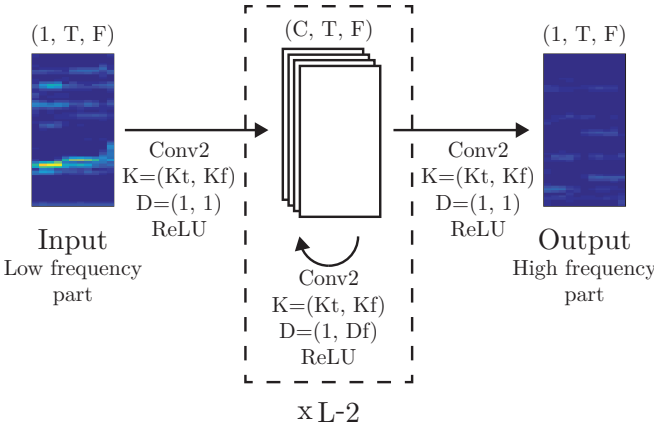

Fig. 1. Proposed deep convolutional neural network architecture for bandwidth extension.

The model is trained using the mean squared error (MSE) loss function. Optimization is performed using the Adam [16] algorithm with minibatches of 64 examples and a learning rate of 0.001.

Reverting to the time domain requires the estimation of the phase information for higher frequencies whose magnitude have been predicted. Several methods proposed in the literature are considered. For evaluation purposes, the phase can be the original phase, termed *oracle* in the following, the low frequency phase as proposed in [13] and a flipped version of the low frequency phase as proposed in [3], termed *mirror* in the following. The latter gave consistently better results, so only the results achieved using this method are reported.

3. EXPERIMENTAL PROTOCOL

3.1. Datasets

Two datasets are considered in this study to evaluate the performance of the proposed model, the *medley-solos-db* dataset [17], and the *gtzan* dataset [18]. The first has 21572 monophonic audio clips of about 3 seconds, for a total duration of about 18 hours. The second comprises 100 polyphonic pop songs of 30 seconds each for each of the 10 musical genres represented, for a total duration of about 8 hours. The *medley-solos-db* dataset is already split into "training", and "test" datasets, they are considered as is in this study. The *gtzan* dataset is split in a train set and a test set using the following procedure. For each genre, the first 70 songs are put in the train set and the remaining 30 in the test set.

For both datasets, the audio data is resampled to 8kHz and converted to spectral data using a short-term Fourier transform (STFT) with frame size of 256 samples, hop size of 128 samples and a Hann window. Extracts are further split into "textures" of 10 frames, processed as individual examples.

The resulting spectra are split into 2 parts, the low frequency one, that serves as input to the different models and

the high frequency one that serves as reference for training the models and computing the evaluation metrics.

3.2. Metrics

Three metrics are considered. The first metric is the average loss on the test set used to get a global understanding of the behavior of the predictors in the spectral domain. The second metric is the average signal-to-reconstruction ratio (SRR), computed in the time domain. Complex spectral values are computed for each batch of textures using the original phase for the low frequency part and one of several estimates for the high frequency part. The time domain signal is obtained by performing an inverse STFT on the complex spectra with the *oracle* phase and the *mirror* phase.

As it will be discussed in Section 4, loss of phase correspondence between the predicted and reference complex spectra lead to the SRR being no longer relevant as a performance metric. Ultimately, subjective evaluation by means of listening tests performed by humans is desirable. However, setting up those tests is time consuming and considering the whole datasets as stimuli is impractical. This issue is left for future research.

A good alternative is the use of objective metrics based on perceptual models that have been proposed and used with success in many audio signal processing tasks, such as PEMO-Q [19] which introduces perceptual similarity measures based on a psychoacoustically validated, quantitative model of the peripheral auditory processing. In this paper, the PSM_t metric is considered.

3.3. Baselines

Three baselines are considered. The first is a simplified version of the SBR technique, where the high frequency part is simply the low frequency part whose amplitudes are scaled. This baseline requires some side information, that is the amplitude scaling factor defined as the ratio of average amplitude between the high frequency part and the low frequency one, computed for each texture.

The second and the third ones are reimplementations of the deep architectures presented in [13]. The *cnn bottleneck* encoder part consists in two convolutional layers with filters of size $(1, F)$ and $(T, 1)$ that independently summarize information in the frequency and time dimensions of the input data respectively. A fully connected layer is then applied to extract a code of 64. The decoder part mirrors these processing steps to recover a spectrum with the original input size. This model is implemented as described in [13], though with different filter sizes as $F = 128$ and $T = 10$ to match the data format considered in this study. The *cnn stride2* autoencoder adopts a different strategy where frequency patterns information is extracted at multiple scales using strided convolutions. The first four layers operate independently on each time frame, and are followed by two convolutional layers with square kernels. The number of channels increases linearly with each additional layer. In this study, the first two layers are removed to account for the difference in the size of the frequency dimension in considered spectra.

Table 1. Spectral loss on the testing set for the proposed architecture with $L = 7$ and $C = 64$.

K	dilation (D)	*medley-solos-db*	*gtzan*
13	1	0.237 $\pm$0.014	0.109 $\pm$0.020
17	1	0.226 $\pm$0.016	0.105 $\pm$0.021
13	2	0.226 $\pm$0.017	0.102 $\pm$0.021
17	2	0.218 $\pm$0.019	0.102 $\pm$0.022

Table 2. Spectral loss on the testing set for the proposed architecture with $K = 17$ and $C = 64$.

L	dilation (D)	*medley-solos-db*	*gtzan*
5	1	0.240 $\pm$0.013	0.110 $\pm$0.017
6	1	0.231 $\pm$0.016	0.107 $\pm$0.021
7	1	0.226 $\pm$0.016	0.105 $\pm$0.021
5	2	0.228 $\pm$0.015	0.102 $\pm$0.019
6	2	0.225 $\pm$0.019	0.102 $\pm$0.022
7	2	0.218 $\pm$0.019	0.102 $\pm$0.022

Additionally, two anchors are considered: the *oracle* predictor that "predicts" the actual high frequency spectral pattern and the *null* predictor that outputs a vector of zeros.

4. EXPERIMENTS

The experiments reported here are conducted in order to study two main design issues for the task at hand: 1) the impact of the dilation on the other meta parameters, and 2) the impact of the compression used in the design of autoencoders.

The dilation parameter D is a very interesting feature of convolutional networks as it allows us to expand the receptive field without increasing the model complexity. The other parameters that control its size are K the size of the filters, and L the number of layers. As can be seen on Table 1, $D = 2$ allows us to reduce the loss and also to reduce the gain achieved by increasing K. The effect is more important on the *gtzan* dataset. Table 2 shows the same effect of D on L.

Increasing further the dilation factor to $D = 3$ is not beneficial, as the receptive field is sufficiently large with $D = 2$. Indeed, with $D = 2$, $K = 17$, $L = 6$, it lead to a receptive field size of 89 where the input and output of the predictor is of size 64. This setting with $C = 64$ is selected for the remaining of the experiments.

Compared to the two encoder-decoder baselines, the proposed approach compares favorably in terms of spectral loss, see Table 3. An advantage of considering spectral bandwidth extension as task is the ease of fine grain performance analysis. Contrary to classification pipelines, the input and the output of the network are equivalent in terms of physical interpretation. It allows us to visually interpret the behavior

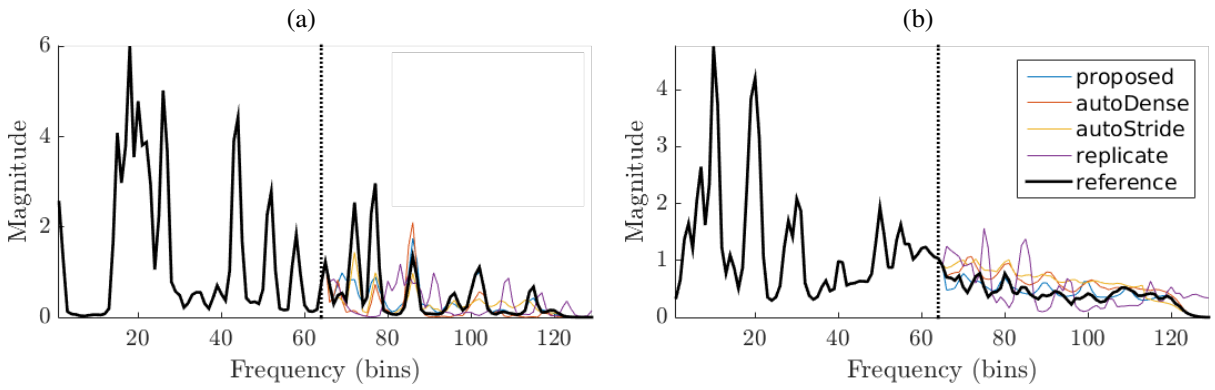

Fig. 2. Examples of predictions for (a) the *medley-solos-db* dataset and (b) the *gtzan* dataset. The proposed model handles correctly the harmonic structures (a) and the average magnitude of more complex spectral shapes (b).

Table 3. Spectral loss over the test set.

method	*medley-solos-db*	*gtzan*
proposed	0.225 $\pm$0.019	0.102 $\pm$0.022
cnn bottleneck	0.241 $\pm$0.011	0.107 $\pm$0.017
cnn stride2	0.228 $\pm$0.013	0.110 $\pm$0.017

of the predictors. As can be seen on Figure 2, the proposed model handles correctly the harmonic structures and the average magnitude of more complex spectral shapes.

In order to evaluate the proposed approach in the time domain, the SRR is considered. As can be seen on Table 4, there is a direct inverse correlation between the spectral loss and the SRR when considering the oracle phase. Considering the mirror phase strongly reduces the phase correlation between the reference and the estimate leading to low SRR even for the oracle magnitude estimator. Though, informal listening test shows that considering the oracle estimator using the mirror phase estimate is perceptually more pleasing than the null predictor. We conclude that the sensitivity of the SRR to phase shift reduces its usefulness and turn to an objective perceptual measure to better assess the performance of the predictors.

As can be seen on Table 5, the proposed approach improves over the baselines in terms of PSM_t (the higher the better), except for the null baseline on the *medley-solosdb*. Whereas the autoencoder architectures appears to perform similarly in terms of SRR, they have different behavior when considering the PSM_t metric, the stride being more effective on monophonic music and the dense being more adapted to polyphonic music.

5. DISCUSSION

The benefit of considering a dilated fully convolutional neural network for the task of predicting the high frequency part of the magnitude spectra given the low frequency part has been studied. The proposed architecture has been thoroughly evaluated in terms of several metrics, from the optimized loss to an objective perceptual measure.

Table 4. SRR achieved by the different methods on the *medley-solosdb* and the *gtzan* datasets using the *mirror* and *oracle* phase estimates.

	medley-solos-db		*gtzan*	
method	mirror	oracle	mirror	oracle
null	12.9 $\pm$3	12.9 $\pm$3	12.9 $\pm$3	12.9 $\pm$3
replicate	10.7 $\pm$3	11.6 $\pm$3	10.8 $\pm$3	13.3 $\pm$3
cnn bottleneck	10.6 $\pm$2	13.3 $\pm$3	11.1 $\pm$3	15.6 $\pm$3
cnn stride2	11.2 $\pm$3	13.9 $\pm$3	11.2 $\pm$3	15.6 $\pm$3
proposed	11.0 $\pm$2	14.3 $\pm$3	11.3 $\pm$3	16.3 $\pm$3
oracle	10.5 $\pm$3	∞	10.5 $\pm$3	∞

Table 5. PSM_t in % using the mirror phase estimate.

method	*medley-solos-db*	*gtzan*
null	90.9 $\pm$1.6	89.7 $\pm$2.3
replicate	86.5 $\pm$1.4	88.4 $\pm$1.8
cnn bottleneck	86.6 $\pm$1.5	91.3 $\pm$1.8
cnn stride2	86.9 $\pm$1.3	90.4 $\pm$2.1
proposed	87.5 $\pm$1.4	91.5 $\pm$1.8
oracle	97.2 $\pm$0.3	97.2 $\pm$0.6

Considering the Fourier spectrum as input has several drawbacks. Time/Frequency resolution tradeoffs and the necessity for phase estimation inherently limit the potential of the proposed approach. Future work will consider more advanced phase estimators and study the influence of mismatch between training and testing audio material.

We also believe that considering this task with varied learning and testing sets, *i.e.* predicting the high frequency spectra of a 'pop' song using network trained on 'country' songs may be of creative interest, to explore the yet to be defined notion of musical style transfer [20].

6. REFERENCES

[1] Erik Larsen and Ronald M Aarts, *Audio bandwidth extension: application of psychoacoustics, signal processing and loudspeaker design*, John Wiley & Sons, 2005.

[2] Johannes Abel and Tim Fingscheidt, "Artificial speech bandwidth extension using deep neural networks for wideband spectral envelope estimation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 1, pp. 71–83, 2017.

[3] Kehuang Li and Chin-Hui Lee, "A deep neural network approach to speech bandwidth expansion," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 4395–4399.

[4] Zhen-Hua Ling, Yang Ai, Yu Gu, and Li-Rong Dai, "Waveform modeling and generation using hierarchical recurrent neural networks for speech bandwidth extension," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 5, pp. 883–894, 2018.

[5] Archit Gupta, Brendan Shillingford, Yannis Assael, and Thomas C. Walters, "Speech bandwidth extension with wavenet," 2019.

[6] Martin Dietz, Lars Liljeryd, Kristofer Kjorling, and Oliver Kunz, "Spectral band replication, a novel approach in audio coding," in *Audio Engineering Society Convention 112*. Audio Engineering Society, 2002.

[7] Per Ekstrand, "Bandwidth extension of audio signals by spectral band replication," in *in Proceedings of the 1st IEEE Benelux Workshop on Model Based Processing and Coding of Audio (MPCA'02*. Citeseer, 2002.

[8] A Ehret, XD Pan, M Schug, H Hoerich, WM Ren, XM Zhu, and F Henn, "Audio coding technology of exac," in *Proceedings of 2004 International Symposium on Intelligent Multimedia, Video and Speech Processing, 2004.* IEEE, 2004, pp. 290–293.

[9] Tobias Friedrich and Gerald Schuller, "Spectral band replication tool for very low delay audio coding applications," in *2007 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*. IEEE, 2007, pp. 199–202.

[10] James L Flanagan and RM Golden, "Phase vocoder," *Bell System Technical Journal*, vol. 45, no. 9, pp. 1493–1509, 1966.

[11] Frederik Nagel and Sascha Disch, "A harmonic bandwidth extension method for audio codecs," in *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2009, pp. 145–148.

[12] Dennis L Sun and Rahul Mazumder, "Non-negative matrix completion for bandwidth extension: A convex optimization approach," in *2013 IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*. IEEE, 2013, pp. 1–6.

[13] M. Miron and M. E. P. Davies, "High frequency magnitude spectrogram reconstruction for music mixtures using convolutional autoencoders," in *In Proc. of the 21st Int. Conference on Digital Audio Effects (DAFx-18)*. IEEE, 2018, pp. 173–180.

[14] Jonathan Long, Evan Shelhamer, and Trevor Darrell, "Fully convolutional networks for semantic segmentation," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2015, pp. 3431–3440.

[15] Fisher Yu and Vladlen Koltun, "Multi-scale context aggregation by dilated convolutions," in *4th International Conference on Learning Representations (ICLR)*, 2016.

[16] Diederik Kingma and Jimmy Ba, "Adam: a method for stochastic optimization," in *3rd International Conference on Learning Representations (ICLR)*, 2015.

[17] Vincent Lostanlen and Carmine-Emanuele Cella, "Deep convolutional networks on the pitch spiral for musical instrument recognition," in *In Proc. of the International Society for Music Information Retrieval Conference (ISMIR)*, 2016.

[18] George Tzanetakis and Perry Cook, "Musical genre classification of audio signals," *IEEE Transactions on speech and audio processing*, vol. 10, no. 5, pp. 293–302, 2002.

[19] Rainer Huber and Birger Kollmeier, "Pemo-q—a new method for objective audio quality assessment using a model of auditory perception," *IEEE Transactions on audio, speech, and language processing*, vol. 14, no. 6, pp. 1902–1911, 2006.

[20] Shuqi Dai, Zheng Zhang, and Gus G Xia, "Music style transfer: A position paper," *arXiv preprint arXiv:1803.06841*, 2018.