JOINT DENOISING, RECONSTRUCTION AND UPSAMPLING OF LOW-BITRATE SPEECH USING DEEP CONVOLUTIONAL DENOISING AUTOENCODERS

Paweł Tomasik[⋆] *Stanisław A. Raczyński*[⋆†]

[⋆] PICTEC, ul. Polanki 12, Gdańsk, Poland, name.surname@pictec.eu
[†] Gdańsk University of Technology, ul. Narutowicza 11/12, Gdańsk, Poland

ABSTRACT

Speech–especially noisy speech–encoded with lossy low-bitrate compression, lacks the higher frequency components and is distorted, which reduces its perceived quality and intelligibility, both for humans and ASR. The task of bandwidth expansion focuses only on reconstruction of higher bands from low-sample-rate signal, while speech enhancement focuses only on removing the noise. We argue that improvement in intelligibility under conditions of noise, low bitrate and transportation loss requires the entire signal to be reconstructed. In this paper, evaluation is done for application of dense and convolutional artificial neural networks to this problem. Several different architectures are compared in noiseless and noisy scenarios. Across experiments, convolutional network with dense heads (ConvDense) achieves the best results.

Index Terms— speech reconstruction and postfiltering, bandwidth expansion, speech denoising, denoising autoencoders, convolutional neural networks

1. INTRODUCTION

Mobile phone networks rely on predictive coding schemes to encode speech signals to adapt to the network bandwidth, which are often very limited. Depending on the location the quality of connection may be compromised. Since the compression ratio in narrowband (NB) codecs is significant, quality depends on the similarity between the transmitted sound and the model used to compress it.

Telephone sound bandwidth is commonly limited to 3400 Hz, which does not contain all speech information. Especially consonant sounds contain high amount of information in high frequency (HF) bands. While this alone is comprehensible for human listeners, AI methods such as automatic speech recognition or speaker recognition [1] may be hindered if the upper band signal is not present.

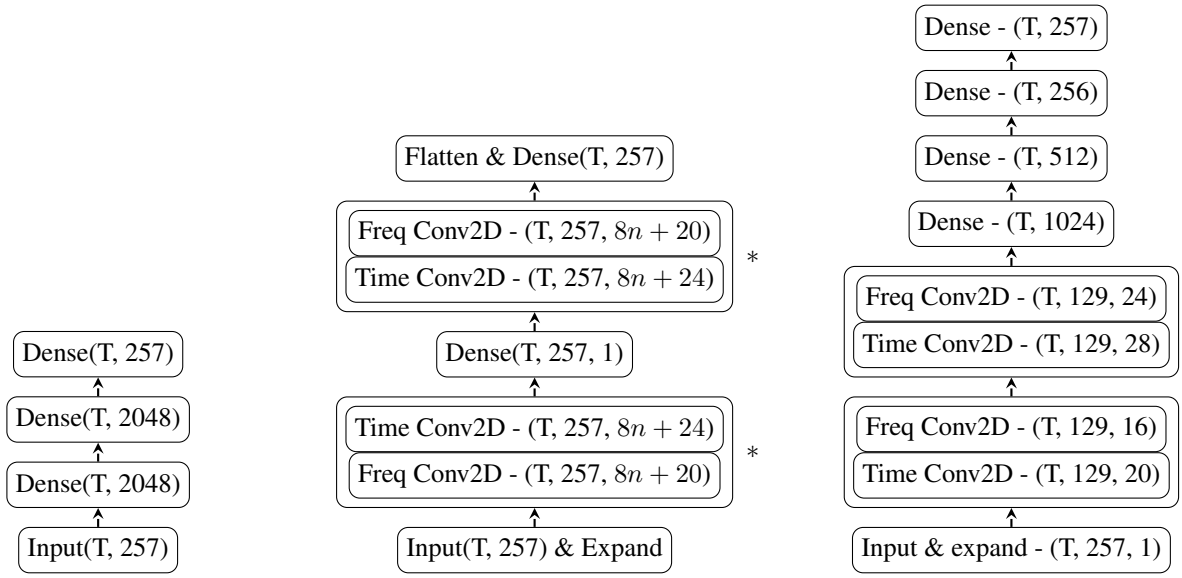
Common telephony codecs (GSM-FR, GSM-EFR[2], AMR[3]) are based on ACELP approach, which is based on the speech production model. Feasibility of these methods relies on the assumption that the sound in question is composed of speech, any background noise may cause the encoder to miscalculate prediction coefficients, further degrading the quality and intelligibility.

The single problem of reconstructing the signal spectrum above 3400 Hz of the narrow-band signal is called bandwidth expansion (BWE) and multiple attempts at providing solutions have been made, including *inter alia* decomposition to spectral shape and excitation [4], Gaussian mixture models [5] and hidden Markov models [6]. Recently, artificial neural network (ANN) methods started being used more commonly, either feed-forward [7], recurrent [8, 4], generative adversarial networks [9] or autoencoders [8, 10].

While most of the work focuses on MSE minimization of magnitude STFT or its logarithm, other representations are used, which may be better suited to representing speech. One particular study estimated cepstra of higher bands (HBs) [11]. Cepstra are also used as a regularization parameter for DNN approach to reduce the mismatch between higher and lower bands [12]. Speech may be described by relatively simple production model, with voiced parts being composed primarily of harmonics. Therefore, authors believe that BWE algorithms should focus on proper reconstruction of harmonic structure.

Most of the cited approaches to BWE deal with clean speech (notable exception is [13]), and usually ignore the distortion introduced by the codec itself, which may be significant in bad conditions. For this reason our approach is closer to postfiltering methods that are used in speech synthesis [14] or compression [15, 16].

In the field of speech denoising, DAEs [17, 18] and convolutional DAEs (CDAEs) [19] are also often used. Convolutional layers are used both in time-frequency, temporal [20] and cepstral domains [16]. More recently, neural models for enhancement of coded speech were devised [16], however they maintained bandwidth and resolution of telephony speech. In this paper it is investigated how convolutional layers and CDAEs perform in improving quality of encoded speech by not only reconstructing the signal in the presence of additive noise, but by joint reconstruction, upsampling and denoising of the speech signal.



(a) Feedforward network (baseline) (b) Convolutional autoencoder (c) Convolutional dense network

Figure 1: Model architectures used in the paper. Asterisk is indicated to use repetition of layer. Depths of 2 and 4 consecutive modules were used. In scenario, where LB and HB were modelled separately, two networks with half the features or channels were used.

2. APPROACH

The goal of proposed solution is to reconstruct the original signal from speech compressed with a linear predictive narrowband codec (either GSM-FR or AMB-NB) which can be recorded in additive noise of varying SNR. The problem consists not only of bandwidth extension, but also involved reconstruction of speech distorted by the compression and removal of background noise. Two scenarios were considered: one corresponding to a simple bandwidth expansion, and one modeling adverse conditions: limited codec bitrate and background noise hindering the quality of signal.

In this work we compare effectiveness of several different configurations of neural networks in the aforementioned task. Authors have experimented with several different variations, which are not included in this paper (The results as well as sample recordings from this paper may be found at `https://research.pictec.eu`) and identified three key models to compare: a baseline model consisting of several feed-forward layers as proposed in [7], a convolutional (denoising) autoencoder (CDAE) and a hybrid approach using convolutional encoder and feed-forward decoder, named hereafter ConvDense. We present the architectures in figure 1.

In each network with convolutional layers, separate filter for time axis and frequency axis is used. Preliminary experiments using two-layer and four-layer with either single 5x3 convolution kernel or separated 1x3 and 5x1 convolutions have shown preference for separated axes (refer to figure 2).

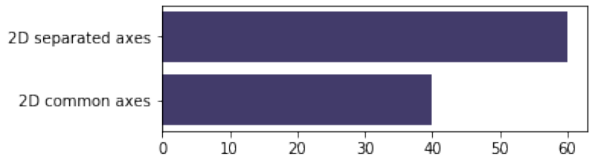


Figure 2: Preferences between separate and common convolution kernels, measured by PESQ comparison between two CDAE models

Rationale for testing ConvDense architecture was related to the detail level in reconstruction of the harmonic frequencies. Figure 3 shows difference between harmonic frequencies in 2kHz-4kHz band between CDAE and ConvDense. While there is no theoretical guarantee one network is better than the other, authors wanted to verify whether the initial observation translated to better quality of reconstructed recordings.

There are several important questions when it comes to choosing a proper network architecture for this task: first of all, whether lower (LB) and higher bands (HB) should be modelled by separate networks or using common representation and layers; how much does the reconstruction of the LB improve speech quality in absence of noise and how much does BWE improve the quality as opposed to using only denoising methods.

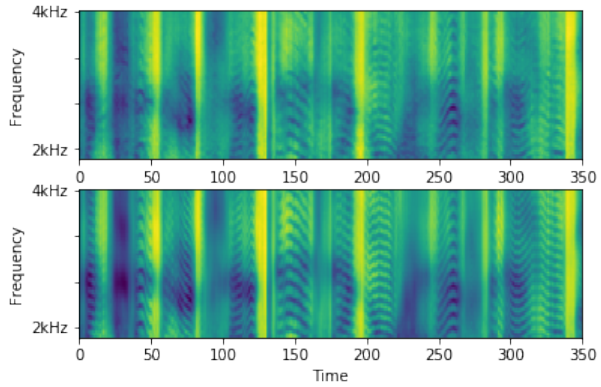


Figure 3: Reconstruction of the recording by CDAE(upper) and ConvDense(lower) in the 2kHz-4kHz frequency range

3. EXPERIMENTS

In order to answer the questions, a series of models was created to compare the corresponding architectures in each of the conditions (noise and codec). In each task we conducted experiments on several models using the same conditions. Optimization was done using Adam optimizer with gradient clipping, restarting the optimizer with lower learning rate after validation loss stopped falling for predetermined number of epochs. Each schedule consisted of four such steps.

We tested two variants of each model, with single series layers for all frequencies or two parallel networks for each lower and upper band. In case of parallel layers, the size of intermediate tensors was halved, resulting in lower amount of parameters.

Librispeech was used as the training dataset [21]. It contains recordings of varying quality, but generally in quiet conditions and from multiple speakers. We randomly selected a subset of recordings used in each training experiment and each recording was trimmed or padded to 10 seconds. The training recordings were the same across models. Total number of training samples was 9000, which without padding would be more than 20 hours of speech. This is of the same order of magnitude as corpora found in related works in the field.

For noise, NOISEX-92 database was used. Noise was mixed additively with the speech signal before compression at random gain. Gain was sampled from a uniform distribution bounded by the gain corresponding to 0db SNR. This was done to ensure models would operate in varying conditions. Noise recordings were resampled to rate of 16kHz.

In every scenario the recordings were converted encoded using a SOX implementation of a codec and decoded back to raw audio. Preprocessed and clean recordings are then converted into respectively 256-point and 512-point log-power STFT with 75% overlap and a Hamming windowing function. Data is normalized to [0, 1] range. Models are trained using MSE criterion and the resulting recordings are resynthesized using original phase padded with zeros.[1]

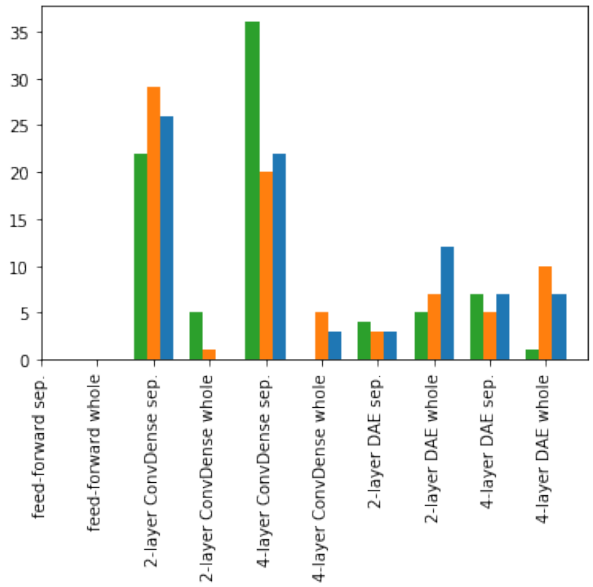


Figure 4: Percentage of best results across the models from the same codec in noisy conditions, from left to right: green - GSM full rate; orange - AMR-NB 5kbit/s; blue - AMR-NB full rate

For evaluation of our approaches we used log power MSE and the perceptual evaluation of speech quality (PESQ), in the wideband *full reference* (FB) mode. We used the reference PESQ implementation and the original wideband clean recordings as a reference. The evaluation was conducted on 80 randomly chosen recordings that had not been used neither in training nor validation. The testing recordings were the same for each model in given scenario.

4. RESULTS

First experiment was conducted on clean recordings, with results listed in table 1. In each scenario, the best reconstruction was performed by ConvDense model. Improvement of full spectrum reconstruction was small, but present in two high-quality codecs, which was consistent with our expectation. However, for low quality AMR codec the models were underperforming in comparison to BWE without reconstruction.

[1] Better methods for phase reconstruction exists, even simply mirroring the phase of LB, but authors think that the results should be consistent nevertheless as UB conveys smaller amount of speech information than LB

In most of the cases, networks with common representation achieved better results than the separated ones, this is however not true for all of the cases, especially for the best result in *GSM separate*.

Second experiment was conducted on noisy recording and its results are listed in table 2. Preferences between different models in each scenario are shown on figure 4. Networks with separate parts for modelling lower and upper band performed better, with best result in each codec obtained by separate networks. In each of those cases, the ConvDense architecture was shown to perform better, indeed as shown on fig. 4. Approx. half of test recordings were reconstructed best by one of ConvDense variants. It is important to note that they were not clearly dominant as sizeable portion of recordings were best reconstructed by autoencoders. Feedforward networks were outclassed by networks containing convolutional layers, which is also confirmed is summary results.

In each noisy scenario, BWE didn't improve the quality and denoising was shown to be more important to the quality of the speech. However, bandwidth expansion improves the quality of the speech in each case, even as much as 0.5 PESQ point in the *GSM separate* scenario. Each network with convolutional layers improved the quality of the signal significantly as opposed to best models used for denoising only.

5. CONCLUSIONS

In this paper we investigated three different architectures and their variations for joint denoising, reconstruction and upsampling. From proposed architectures, ConvDense exhibited the best behaviour, showing best results in ten out of twelve scenarios. However, it cannot be said that ConvDense is a decisively better architecture, as some of the noisy recordings were reconstructed better by CDAE network.

Using common features for the reconstruction of both LB and UB was shown to degrade performance on noisy recordings, which is against the assumption that using single network would leverage common representation. In most cases, it improved the reconstruction on clean recordings.

It was shown that upsampling with denoising show significantly better results than denoising alone. Improvement of quality due to reconstruction of lower band on clean recordings is small, the models did not consistently show better results as some reconstructions performed worse than supplying original LB, however properly tuned network outperforms plain BWE.

5.1. Future directions

MSE is not an optimal loss function, and other, perceptually motivated ones (such as MFCC) could be used instead, either alone or as an additional term. GAN networks are also attractive as a method of obtaining realistic sound and could be used in this scenario. More explanation could be found on the reason why the common representation doesn't help the denoising BWE models to achieve better results.

PESQ	**AMR joint**	**AMR 5kb/s joint**	**GSM joint**	**AMR separate**	**AMR 5kb/s sep.**	**GSM separate**
Only expansion	2.756 ± 0.118	*NA*	2.813 ± 0.142	2.748 ± 0.122	$\mathbf{2.705 \pm 0.122}$	2.825 ± 0.127
Feed-forward	2.260 ± 0.128	2.484 ± 0.104	2.918 ± 0.114	2.845 ± 0.127	$1.939 \pm 0.0.89$	2.446 ± 0.116
2-layer ConvDAE	2.824 ± 0.118	2.553 ± 0.114	2.922 ± 0.129	2.763 ± 0.158	2.467 ± 0.122	2.825 ± 0.173
4-layer ConvDAE	2.873 ± 0.151	2.624 ± 0.114	2.810 ± 0.147	2.827 ± 0.146	2.481 ± 0.119	2.374 ± 0.137
2-layer ConvDense	$\mathbf{2.896 \pm 0.122}$	2.693 ± 0.118	$\mathbf{2.986 \pm 0.131}$	$\mathbf{2.883 \pm 0.132}$	2.532 ± 0.120	$\mathbf{3.012 \pm 0.122}$
4-layer ConvDense	2.353 ± 0.127	$\mathbf{2.698 \pm 0.128}$	*NA*	2.518 ± 0.127	2.255 ± 0.110	2.627 ± 0.129

Table 1: PESQ measurements of reconstructed clean recordings; expansion only uses the best model from a selection of models trained specifically to reconstruct upper band; best result in each scenario is highlighted

PESQ	**AMR joint**	**AMR 5kb/s joint**	**GSM joint**	**AMR separate**	**AMR 5kb/s sep.**	**GSM separate**
Expansion only	1.140 ± 0.054	1.108 ± 0.037	1.139 ± 0.056	1.104 ± 0.036	1.143 ± 0.060	1.128 ± 0.054
Denoising only	1.706 ± 0.087	1.685 ± 0.087	1.724 ± 0.094	1.725 ± 0.092	1.720 ± 0.091	1.717 ± 0.093
Dense	1.751 ± 0.082	1.661 ± 0.078	1.627 ± 0.069	1.804 ± 0.088	1.712 ± 0.085	1.715 ± 0.081
2-layer ConvDAE	2.118 ± 0.118	1.877 ± 0.096	2.059 ± 0.101	2.095 ± 0.122	1.794 ± 0.098	2.049 ± 0.103
4-layer ConvDAE	$\mathbf{2.119 \pm 0.111}$	$\mathbf{1.910 \pm 0.098}$	1.975 ± 0.087	2.094 ± 0.119	1.823 ± 0.101	2.023 ± 0.102
2-layer ConvDense	2.031 ± 0.103	1.850 ± 0.094	$\mathbf{2.075 \pm 0.101}$	$\mathbf{2.228 \pm 0.129}$	$\mathbf{1.975 \pm 0.103}$	2.197 ± 0.114
4-layer ConvDense	2.042 ± 0.107	1.865 ± 0.096	1.999 ± 0.095	2.203 ± 0.126	1.930 ± 0.104	$\mathbf{2.232 \pm 0.120}$

Table 2: PESQ measurements of reconstructed recordings in noise with $[0, \infty)$ dB SNR; expansion and denoising use the best model with proper modifications; best result in each scenario is highlighted

6. REFERENCES

[1] P. S. Nidadavolu, C.-I Lai, J. Villalba, and N. Dehak, "Investigation on bandwidth extension for speaker recognition," in *Interspeech 2018, 19th Annual Conference of the International Speech Communication Association*, 2018, pp. 1111–1115.

[2] K. Jarvinen, J. Vainio, P. Kapanen, T. Honkanen, P. Haavisto, R. Salami, C. Laflamme, and J. . Adoul, "GSM enhanced full rate speech codec," in *1997 IEEE International Conference on Acoustics, Speech, and Signal Processing*, April 1997, vol. 2, pp. 771–774 vol.2.

[3] A. Uvliden, S. Bruhn, and R. Hagen, "Adaptive multirate. a speech service adapted to cellular radio network quality," in *Conference Record of Thirty-Second Asilomar Conference on Signals, Systems and Computers (Cat. No.98CH36284)*, Nov 1998, vol. 1, pp. 343–347 vol.1.

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

[5] M. S. Ganesh, B. Patnaik, and M. Karthik, "Narrowband speech signal bandwidth extension for intelligible speech communication," in *IEEE International Conference on Intelligent Techniques in Control, Optimization and Signal Processing (INCOS)*, 2017, pp. 1–5.

[6] J. Han, G. J. Mysore, and B. Pardo, "Language informed bandwidth expansion," in *2012 IEEE International Workshop on Machine Learning for Signal Processing*, 2012, pp. 1–6.

[7] K. Li and C. H. Lee, "A deep neural network approach to speech bandwidth expansion," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 4395–4399.

[8] Y. Gu, Z.-H. Ling, and L.-R. Dai, "Speech bandwidth extension using bottleneck features and deep recurrent neural networks," in *Interspeech*, 2016, pp. 297–301.

[9] S. Li, S. Villette, P. Ramadas, and D. J. Sinder, "Speech bandwidth extension using generative adversarial networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5029–5033.

[10] V. Kuleshov, S. Zayd Enam, and S. Ermon, "Audio super-resolution using neural nets," in *International Conference on Machine Learning*, 2017.

[11] J. Abel, M. Strake, and T. Fingscheidt, "A simple cepstral domain DNN approach to artificial speech bandwidth extension," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5469–5473.

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

[13] B. Lee, K. Noh, J. Chang, K. Choo, and E. Oh, "Sequential deep neural networks ensemble for speech bandwidth extension," *IEEE Access*, vol. 6, pp. 27039–27047, 2018.

[14] M. Coto-Jiménez and J. G. Close, "LSTM deep neural networks postfiltering for improving the quality of synthetic voices," *CoRR*, vol. abs/1602.02656, 2016.

[15] Z. Zhao, H. Liu, and T. Fingscheidt, "Convolutional Neural Networks to Enhance Coded Speech," *ArXiv e-prints*, June 2018.

[16] Ziyue Zhao, Huijun Liu, and Tim Fingscheidt, "Convolutional neural networksto enhance coded speech," January 2019.

[17] S. Matsuda C. Hori X. Lu, Y. Tsao, "Ensemble modeling of denoising autoencoder for speech spectrum restoration," in *Interspeech*, 2014, pp. 885–889.

[18] M. Sun, X. Zhang, H. Van hamme, and T. F. Zheng, "Unseen noise estimation using separable deep auto encoder for speech enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 1, pp. 93–104, Jan 2016.

[19] T. Kounovsky and J. Malek, "Single channel speech enhancement using convolutional neural network," in *2017 IEEE International Workshop of Electronics, Control, Measurement, Signals and their Application to Mechatronics (ECMSM)*, 2017, pp. 1–5.

[20] DeLiang Wang and Jitong Chen, "Supervised speech separation based on deep learning: An overview," *ArXiv e-prints*, January 2019.

[21] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2015, pp. 5206–5210.