BLIND BANDWIDTH EXTENSION BASED ON CONVOLUTIONAL AND RECURRENT DEEP NEURAL NETWORKS

Konstantin Schmidt, Bernd Edler

International Audio Laboratories Erlangen
Am Wolfsmantel 33
91058 Erlangen, Germany

ABSTRACT

A blind bandwidth extension (BBWE) expands the bandwidth of telephone speech which often is limited to 0.2 to 3.4 kHz. The advantage is an increased perceived quality as well as an increased intelligibility. This work presents a BBWE similar to state-of-the-art bandwidth extensions like *Intelligent Gap Filling* with the difference that all processing is done in the decoder without the need of transmitting extra bits. Parameters like spectral envelope are estimated by a regressive Convolutional Deep Neuronal Network (CNN) with long short-term memory (LSTM). The system operates on frames of 20 ms without additional algorithmic delay and can be applied in state-of-the-art speech and audio codecs.

Index Terms— Blind Bandwith Extension, Artificial Bandwidth Extension, Speech Coding, Audio Coding, DNN, regressive DNN, LSTM, CNN

1. INTRODUCTION

Todays most used codec for mobile speech communication is still AMR-NB which encodes only frequencies from 200 to 3400 Hz (usually named *narrowband*, (NB)). The human speech signal though has a much wider bandwidth - especially fricatives often have most of their energy above 4 kHz. Limiting the frequency range of speech will not only sound less pleasant but will also be less intelligible [1, 2].

State-of-the-art audio codecs like EVS [3] are able to code a much wider frequency range of the signal but using these codecs will require a change of the whole communication network including the receiving devices. This is a huge effort and known to last several years. Blind bandwidth extensions (BBWE - also known as artificial bandwidth extension or blind bandwidth expansion) are able to extent the frequency range of a signal without the need of additional bits. They are applied to the decoded signal only and do not need any adaption of the network or the sending device. While being an appealing solution to the problem of limited bandwidth of narrow band codecs lots of systems fail to improve the quality of speech signals. In a joint evaluation of latest bandwidth extensions, only four out of 12 systems managed to improve the perceived quality significantly for all tested languages [4].

Following the source-filter model of speech production most bandwidth extensions (blind or non-blind) have two main building blocks - the generation of an excitation signal and estimation of the vocal tract shape. This is also the approach the presented system follows. A commonly used technique for generating the excitation signal is spectral folding, translation or nonlinear processing. The vocal tract shape is often generated by Gaussian Mixture Models (GMM), Hidden Markov Models (HMM), Neural Networks or Deep Neural Networks (DNN). These models predict the vocal tract shape from features calculated on the speech signal.

In [5] and [6] the excitation signal is generated by spectral folding and the vocal tract filter is realized as all-pole filter in time-domain by an HMM. First a codebook of linear prediction coefficients (LPC) calculated on frames containing the upper band speech signal is created by vector quantization. At decoder-side features are calculated on the decoded speech signal and an HMM is used to model the conditional probability of a codebook entry given the features. The final envelope is the weighted sum of all codebook entries with the probabilities being the weights. In [6] fricative sounds are additionally emphasized by a neural network.

In [7] the excitation signal is also generated by spectral folding and the vocal tract is modeled by a neural network which outputs gains applied to the folded signal in a Mel filterbank domain.

In [8] a DNN is used to predict the spectral envelope of a spectral folded excitation signal (phrased here as imaged phase). The system in [9] also uses the spectral folded excitation signal and shapes the envelope by a DNN comprising LSTM layers. Using several frames of audio as input for the DNN these two systems have an algorithmic delay too high for realtime telecommunication.

A recent approach directly models the missing signal in time-domain [10] with an algorithmic delay of 0 to 32 ms with an architecture similar to WaveNet [11].

The main contribution of this work is a BBWE that operates on frames of 20 ms and exploits the performance of state-of-the-art convolutional and recurrent networks to model the spectral envelope of speech signals. The next section will outline the proposed blind bandwidth extension and section 3 will focus on the deep neural network. Finally the system will be evaluated in section 4 followed by a some conclusions in 5.

2. SYSTEM OVERVIEW

A basic acoustic model of the human speech production process combines a periodic, pulse-like excitation signal (the larynx signal) modulated by a transfer filter determined by the shape of the supralaryngeal vocal tract. Furthermore there are noise-like signals that result form turbulent air flow caused by constriction of the vocal tract or the lips. Based on this model the missing frequency range is extended by extending a spectrally flat excitation signal and then shaping it with an estimate of the vocal tract filter. Figure 1 depicts the proposed system. From the decoded time-domain signal blocks of 20 ms are transformed by a DFT to the frequency domain. The frame increment (hop-size) of adjacent frames is 5 ms. In the frequency domain the signal is upsampled to 16 kHz by zero-padding

and the missing frequency content above 7 kHz is generated in the same way as in bandwidth extensions like *Intelligent Gap Filling* (IGF) or SBR [12, 13]: the lower bins are copied-up to create the missing signal. Since codecs like AMR-NB only code frequencies between 200 and 3400 Hz this signal is not enough to fill the missing range of 8000-3200 = 4800 Hz. Therefore this operation has to be done twice - first time to fill the range of 3400 to 6600 Hz and another time to fill the range of 6600 to 8000 Hz.

This artificial generated signal is too tonal compared to the original excitation signal. A low complex method used in IGF is used to reduce the tonality [14]. The idea here is to divide the signal by its spectral envelope generated by FIR-filtering the power spectrum. This serves two purposes - first the formant structure is removed from the copied signal (this could also be achieved by using the LPC residual), second the ratio of the energy of the harmonics to the noise is lowered. Therefore this signal will sound much more natural.

After an inverse DFT of double the size than the initial DFT the time-domain signal with 16 kHz sampling frequency is generated by overlap-adding blocks with 50% overlap. This time-domain signal with flat excitation signal above 3400 Hz will now be shaped to resemble the formant structure of the original signal. This is done in the frequency domain of a DFT with higher time-resolution operating on blocks of 10 ms. Here the signal in the range of 3400 to 8000 Hz is divided into 5 bands of roughly 1 bark width [15] and each DFT-bin X_i inside band b is scaled by a scaling factor f_b:

$$\hat{X}_i = X_i \sqrt{f_b}. \tag{1}$$

The scaling factor f_b is the ratio of the logarithmic energy estimate L_b and mean energy of the bins i in band b:

$$f_b = \frac{e^{L_b}}{\sum_j |X_j|^2}, \tag{2}$$

where j iterates over all bins inside band b. L_b is calculated by a DNN explained in the next section and is an estimate of the true wide-band energies $\widetilde{L}_b$:

$$\widetilde{L}_b = \log \sum_j |\widetilde{X}_j|^2, \tag{3}$$

which is calculated on the spectrum of the original wide-band signal $\widetilde{X}$.

Finally the scaled spectrum $\hat{X}_i$ is converted to time-domain by an inverse DFT and the output signal is generated by overlap-adding previous frames with 50 % overlap.

3. DNN

The target energy estimate L_b in equation 2 in section 2 scales the spectrum of the synthesized signal to approximate the energy of the original signal. This value is calculated by a DNN. The input to the DNN are concatenated frames of the lower band power spectrum. This is different to state-of-the-art methods where the input are features like *Mel Frequency Cepstral Coefficients*. Instead the first DNN layers are convolutional layers (CNN) followed by LSTM layers and a final fully connected layer with linear activation functions.

CNNs are a variation of multilayer perceptrons inspired by the organization of receptive fields in eyes. A CNN layer is a layer of filter kernels with the kernel coefficients learned during training [16]. CNNs exploit local dependencies much better and with fewer trainable coefficients than fully connected layers. The dimension of the filter kernel is in principle arbitrary but should not exceed the dimension of the input data. Here two-dimensional filter kernels are convolved with the input spectrogram in time and frequency dimension. These filter are able to detect abstract pattern in the signal similar to features like i.a. *spectral centroid* or *Mel Frequency Cepstral Coefficients*.

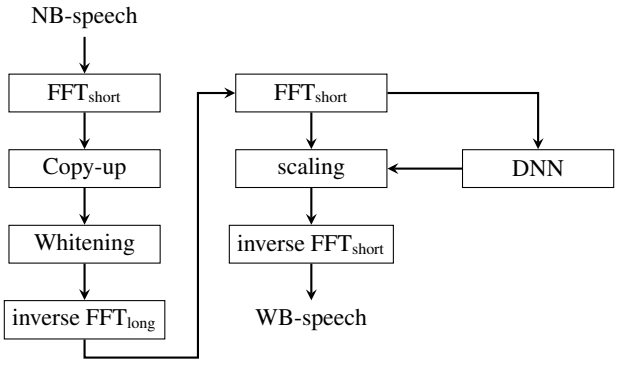

Fig. 1. Block diagram of the proposed system. Input is a narrowband (NB) signal that is extended to a wideband (WB) signal in two stages. In the first 4 blocks the signal is upsampled to 16 kHz and the excitation signal is generated. In the remaining blocks the WB envelope is shaped by a DNN in the frequency domain of a DFT with a higher time-resolution

The convolutional layers are followed by recurrent layers. Recurrent layers are suited to learn longer time-dependencies. There are different types of recurrent layers and here LSTM-layers showed the best performance. LSTMS are able to exploit *short* as well as *long* time structure [17]. Similar but slightly less performance could be achieved with layers of *gated recurrent units* (GRU) [18].

The last layer of the network is a fully connected layer with linear output function. The linear output function allows the network to output unlimited continuous values.

The DNN is trained in a supervised manner by minimizing the difference between the energies of the true wide-band spectrum $\widetilde{L}_b$ and the per iteration estimate L_b. For this a variant of the mini-batch stochastic gradient descent algorithm (SGD) called *Adagrad* [19] was used. Like in standard SGD the networks parameters are iteratively updated until a local minimum of a predefined loss-function is reached but no learning rate has to be tuned by hand.

An important aspect is the definition of the loss function. Since the system will ultimately be judged by humans listeners a perceptual motivated loss is beneficial. Furthermore the training shall be done with deep learning libraries like Keras [20] and for this reason the loss and its derivative must be able to be calculated efficient on CPUs or GPUs. In this work the logarithm in equation 3 implements a coarse loudness model. The advantage of this is that the error function reduces to the euclidian distance. Replacing the logarithm in equation 3 by $()^{\frac{1}{3}}$ has also been tried but informal listening didn't show any benefits.

Another important aspect is the algorithmic delay of the DNN since the presented system should be used in realtime applications. Because the DNN operates on concatenated frames with a frame-increment of one frame the main source of delay comes from the first convolutional layer. In favor of keeping the delay as low as possible the time-dimension of the kernel was set to three - meaning a kernel covers three frames. Since the DNN operates on shorter frames than

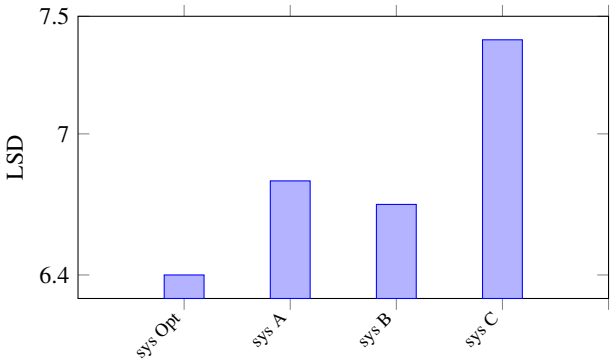

Fig. 2. Comparing the performance of different DNN configurations. System Opt has two convolutional layers (4 kernels) followed by two LSTM layers (16 units each). System A has a single CNN layer (4 kernels) and a single LSTM layer (16 units). System B has no CNN layer but two LSTM layers (32 and 16 units). System C has two CNN layers (4 kernels each)

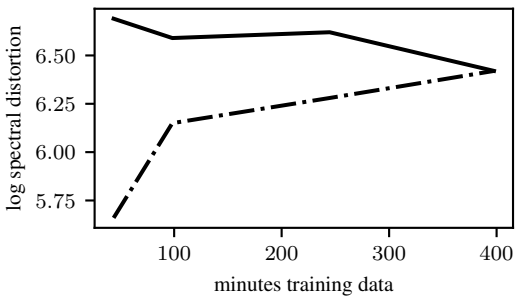

Fig. 3. Error on training set (dashed line) and test set (solid line) dependent on amount of data. With few training data (100 minutes or less) strong overfitting occurs. With a training set of more than 400 minutes overfitting is eliminated

the upsampling and excitation generation in 2 the convolutional layer doesn't add additional algorithmic delay. In frequency direction the kernels cover 250 Hz. Other kernel sizes have been tested but didn't improve the performance.

3.1. Training Data

One important aspect of training a DNN is the versatility of the training set. In order to build a model that is large enough to model the highly non-linear characteristics of the vocal tract the training set needs to be large and contain a vast variety of data - namely different speakers with different languages all of this recorded with different recording gear in different rooms. The 400 minutes long training set has been compiled from several public accessible speech corpora [21] as well as in-house recordings. The training set contains native spoken speech including the following languages: native American English, Arabic, Chinese (Mandarin), Dutch, English (British), Finnish, French, German, Greek, Hungarian, Italian, Japanese, Korean, Polish, Portuguese (Brazilian), Russian, Spanish (Castilian), Swedish. The evaluation set neither contains speaker from the training set nor a recording setup used in the training set and is 8 minutes long.

4. EVALUATION

The presented system was evaluated by objective and subjective tests. First the structure of the network was optimized by maximizing *Logarithmic Spectral Distortion* or LSD. LSD is a well-known measure used in most publications regarding quantization of Linear Prediction Coefficients and correlates well with subjective perception:

$$LSD = \frac{1}{M}\sum_{i=0}^{M-1}\sqrt{\frac{1}{N}\sum_{j=0}^{N-1}(10\log_{10}|X_j| - 10\log_{10}|\widetilde{X}_j|)^2},$$

where $\widetilde{X}$ is the upper band spectrum of the original signal, X is the upper band spectrum of the predicted signal and N is the number of bins in the upper band. M is the number of frames used for the evaluation.

Figure 2 compares the performance of different DNN configurations. The best performing system (Opt) has two convolutional layers with 4 filter per layer, followed by two LSTM layers with 16 units each layer. System A has a single CNN layer with 4 kernels and a single LSTM layer with 16 units. System B has no CNN layer at all but two LSTM layers (32, and 16 units). System C has two CNN layers (4 filter per layer) and no LSTM layer. Here it shows that LSTM layers have the biggest influence on the performance. A system with no LSTM layer performs much worse than a system with LSTM layer. The influence of the convolutional layer on the performance is less - a system without a convolutional layer still performs only 0.5 dB worse than the best system.

Figure 3 shows the influence of the amount of training data on the performance. Small training sets may lead to models that perform very well on the training set but not on unknown data. Here it shows that a training set of 400 and more minutes is enough to create a model with almost no overfitting. Of course this may not be generalized to models with much higher capacity.

Table 1 evaluates the performance of a training and test set mismatch - one being coded with AMR-NB, the other one being uncoded. The left column shows the performance of the DNN trained on speech coded with AMR-NB, the right column shows the performance of a DNN trained on uncoded speech. In the upper row the test set was coded with AMR-NB, in the lower row the test set was uncoded. Apparently a DNN trained on speech coded with AMR-NB performs better in a situation where the system would be applied to uncoded speech than vice versa. In addition AMR-NB degrades the performance of almost half a dB.

	DNN AMR-NB	DNN uncoded
test set AMR-NB	6.4	7.8
test set uncoded	7.5	6.0

Table 1. Performance of the DNN being trained with speech coded with AMR-NB (left column) or with uncoded speech (right column) evaluated on test sets being coded with AMR-NB (upper row) or uncoded (lower row). Performance shown as log spectral distortion (LSD)

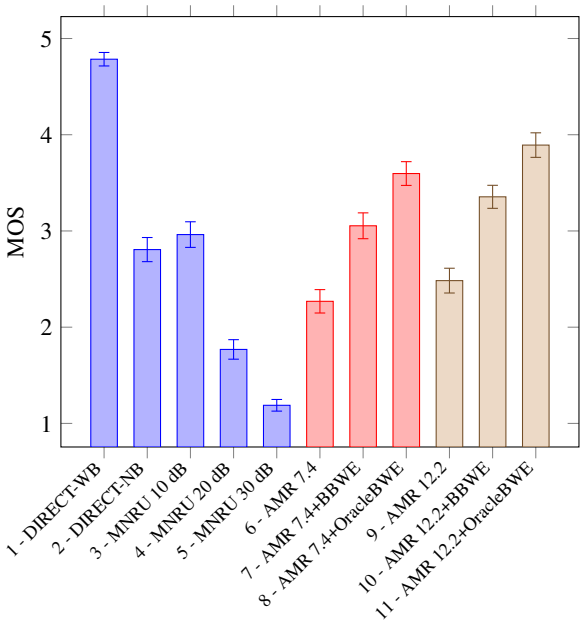

Fig. 4. Results form the ACR listining test displayed as MOS values with 95% confidence intervals. The codecs under test are - from left to right - 1) direct wide-band 2) direct narrow-band 3-5) MNRU 10 - 30 dB noise 6) AMR-NB 7.4 kbps 7) AMR-NB 7.4 kbps with blind bandwidth extension 8) AMR-NB 7.4 kbps with oracle BWE 9) AMR-NB 12.2 kbps 10) AMR-NB 12.2 kbps with BBWE 10) AMR-NB 12.2 kbps with oracle BWE

4.1. Listening Test

Finally the presented system was evaluated with a listening test with the same test method as in [4]. The test is an Absolute Category Rating (ACR) test [22] where a stimulus is presented to a listener without any reference. The listener rates the stimulus on a scale from 1 to 5 (Mean Opinion Score, MOS). 29 unexperienced listeners participated in the test and the test material were 30 recordings of both female and male speech without background noise. Each recording contains a sentence pair and was 8 s long. Each condition was tested with 6 different speech files from 3 female and 3 male speakers. Before the main test started, six speech files of different processing conditions and speakers were presented to the participants in order to accustom them to the range of qualities to be experienced in the test.

The results from the test are presented in figure 4 displayed as average MOS-values with 95 % confidence intervals. The direct WB condition achieved the highest ratings of 4.8 MOS while the direct NB condition achieved 2.8 MOS. Next are the Modulated Noise Reference Units (MNRU) [23] which is speech degraded by modulated noise (sampled at 16 kHz). They serve as quality anchor and make the test comparable to other tests. Finally the results of AMR-NB, AMR-NB with the presented blind bandwidth extension and AMR-NB with an oracle bandwidth extension are shown at two different bitrates - 7.4 kbps and 12.2 kbps. The oracle system differs from the presented system by scaling the spectrum to reach the energy of the original. This is done by replacing the DNN estimate L_b in equation 2 by $\widetilde{L}_b$ calculated on the original WB spectrum. This system is an upper bound of quality a bandwidth extension could reach.

The results show that presented bandwidth extension works well by improving the quality of AMR-NB by 0.8 MOS (7 kbps) to 0.9 MOS (12.2 kbps). The BBWE at 12.2 kbps is also significant better than the direct NB condition. Nevertheless there is still lot of space for improvement as the results from the oracle BWE show.

5. CONCLUSION

A blind bandwidth extension was presented that is able to improve the quality of AMR-NB by 0.8 - 0.9 MOS. It does not add additional algorithmic delay to AMR-NB. The complexity is also moderate so it can be implemented on mobile devices. The system can be easily adopted to different core codecs and reconfigured to different bandwidth settings.

6. REFERENCES

[1] Patrick Bauer, Rosa-Linde Fischer, Martina Bellanova, Henning Puder, and Tim Fingscheidt, "On improving telephone speech intelligibility for hearing impaired persons," in *Proceedings of the 10. ITG Conference on Speech Communication, Braunschweig, Germany, September 26-28, 2012*, 2012, pp. 1–4.

[2] Patrick Bauer, Jennifer Jones, and Tim Fingscheidt, "Impact of hearing impairment on fricative intelligibility for artificially bandwidth-extended telephone speech in noise," in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2013, Vancouver, BC, Canada, May 26-31, 2013*, 2013, pp. 7039–7043.

[3] Stefan Bruhn, Harald Pobloth, Markus Schnell, Bernhard Grill, Jon Gibbs, Lei Miao, Kari Järvinen, Lasse Laaksonen, Noboru Harada, N. Naka, Stéphane Ragot, Stéphane Proust, T. Sanda, Imre Varga, C. Greer, Milan Jelinek, M. Xie, and Paolo Usai, "Standardization of the new 3GPP EVS codec," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2015, South Brisbane, Queensland, Australia, April 19-24, 2015*, 2015, pp. 5703–5707.

[4] Johannes Abel, Magdalena Kaniewska, Cyril Guillaume, Wouter Tirry, Hannu Pulakka, Ville Myllylä, Jari Sjoberg, Paavo Alku, Itai Katsir, David Malah, Israel Cohen, M. A. Tugtekin Turan, Engin Erzin, Thomas Schlien, Peter Vary, Amr H. Nour-Eldin, Peter Kabal, and Tim Fingscheidt, "A subjective listening test of six different artificial bandwidth extension approaches in english, chinese, german, and korean," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2016, Shanghai, China, March 20-25, 2016*, 2016, pp. 5915–5919.

[5] Peter Jax and Peter Vary, "Wideband extension of telephone speech using a hidden markov model," in *2000 IEEE Workshop on Speech Coding. Proceedings.*, 2000, pp. 133–135.

[6] Patrick Bauer, Johannes Abel, and Tim Fingscheidt, "Hmm-based artificial bandwidth extension supported by neural networks," in *14th International Workshop on Acoustic Signal Enhancement, IWAENC 2014, Juan-les-Pins, France, September 8-11, 2014*, 2014, pp. 1–5.

[7] Hannu Pulakka and Paavo Alku, "Bandwidth extension of telephone speech using a neural network and a filter bank implementation for highband mel spectrum," *IEEE Trans. Audio, Speech & Language Processing*, vol. 19, no. 7, pp. 2170–2183, 2011.

[8] 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 2015, South Brisbane, Queensland, Australia, April 19-24, 2015*, 2015, pp. 4395–4399.

[9] Yu Gu, Zhen-Hua Ling, and Li-Rong Dai, "Speech bandwidth extension using bottleneck features and deep recurrent neural networks," in *Interspeech 2016, 17th Annual Conference of the International Speech Communication Association, San Francisco, CA, USA, September 8-12, 2016*, 2016, pp. 297–301.

[10] Yu Gu and Zhen-Hua Ling, "Waveform modeling using stacked dilated convolutional neural networks for speech bandwidth extension," in *Interspeech 2017, 18th Annual Conference of the International Speech Communication Association, Stockholm, Sweden, August 20-24, 2017*, 2017, pp. 1123–1127.

[11] Aäron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew W. Senior, and Koray Kavukcuoglu, "Wavenet: A generative model for raw audio," in *The 9th ISCA Speech Synthesis Workshop, Sunnyvale, CA, USA, 13-15 September 2016*, 2016, p. 125.

[12] Sascha Disch, Andreas Niedermeier, Christian R. Helmrich, Christian Neukam, Konstantin Schmidt, Ralf Geiger, Jérémie Lecomte, Florin Ghido, Frederik Nagel, and Bernd Edler, "Intelligent gap filling in perceptual transform coding of audio," in *Audio Engineering Society Convention 141, Los Angeles*, Sep 2016.

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

[14] Konstantin Schmidt and Christian Neukam, "Low complexity tonality control in the intelligent gap filling tool," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2016, Shanghai, China, March 20-25, 2016*, 2016, pp. 644–648.

[15] Hugo Fastl and Eberhard Zwicker, *Psychoacoustics: Facts and Models*, Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.

[16] Yann Lecun, Léon Bottou, Yoshua Bengio, and Patrick Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov 1998.

[17] Sepp Hochreiter and Jürgen Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[18] Junyoung Chung, Çaglar Gülçehre, KyungHyun Cho, and Yoshua Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *NIPS Deep Learning workshop, Montréal, Canada*, 2014.

[19] John C. Duchi, Elad Hazan, and Yoram Singer, "Adaptive subgradient methods for online learning and stochastic optimization," in *COLT 2010 - The 23rd Conference on Learning Theory, Haifa, Israel, June 27-29, 2010*, 2010, pp. 257–269.

[20] François Chollet et al., "Keras 1.2.2," `https://github.com/fchollet/keras`, 2015.

[21] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2015, South Brisbane, Queensland, Australia, April 19-24, 2015*, 2015, pp. 5206–5210.

[22] ITU-T, "ITU-T recommendation P.800. methods for objective and subjective assessment of quality," 1996.

[23] ITU-T, "ITU-T recommendation P.810. modulated noise reference unit (MNRU)," 1996.