APPROACH TO SPEAKER-GENERALIZED SPECTRAL ENVELOPE ESTIMATION BY DEEP RECURRENT NEURAL NETWORK FOR SPEECH RECONSTRUCTION IN A SPEECH ENHANCEMENT SYSTEM

Stefan Ciba, Mohammed Krini, Amir Rajabi
Aschaffenburg University of Applied Sciences
{first name.last name}@th-ab.de

Abstract: Classical algorithms for *Speech Enhancement* (SE) often show unsatisfactory results in loud or exquisite noise scenarios which can be stationary or transient. Data-driven estimation techniques can outperform classical algorithms by means of machine learning. In this work machine learning and the classical approach are brought together in a *Speech Enhancement System* (SE-System). Furthermore, the source-filter model of speech production is used in such a way, that spectral speech features, namely excitation and envelope, can be estimated separately and subsequently combined. The estimation of the envelope was done by a Deep Recurrent Neural Network (Deep-RNN) as regressive model. The spectral envelope is extracted by an *Infinite Impulse Response-Filter* (IIR-Filter). The Deep-RNN is trained with many speakers and tested with several unseen speakers, to approach speaker generalization. Finally, the estimations as well as the potential of signal improvements, by applying the ideal excitation by the SE-System, are measured and discussed.

1 Introduction

The utilization of single-channel speech enhancement, which uses the *divide-and-conquer* principal to solve the challenging task of SE in two stages is considered in this work. This approach uses source-filter model to split the task into two sub-problems. This sub-problems are then solved by separately improving the source (the excitation signal) and the filter model (the spectral envelope) used for speech reconstruction [1].
Studies on the spectral envelope estimation with repetitive *Long Short-Term Memory* (LSTM) for single speaker and stationary car noise scenarios have been done before [2]. In order to measure the improvement of spectral envelope quality *Log Spectral Distance* (LSD) [3] is used. The improvement of the spectral envelope during speech was 6.42 dB at a *Signal to Noise Ratio* (SNR) of 0 dB. Estimations of cepstral envelope representation showed less improvement compared to the spectral envelope approach. However, this work focuses on the more promising spectral envelope estimation with the further goal of a speaker generalization in a SE-System.
To determine the improvement potential of the reconstruction in a SE-System, the clean excitation is utilized, making the reconstruction ideal concerning the excitation part. The *Short-Time Objective Intelligibility* (STOI), which is highly correlated to envelope quality, is used to measure the improvement of intelligibility [4]. The *Perceptual Objective Listening Quality Analysis* (POLQA) [5] algorithm is used to measure the improvement in listening quality.

2 State of the Art

The utilization of the source-filter model is one of many ways to enhance speech. The envelope and excitation make up the main share of improvement potential since they map a magnitude

spectrum in two feature-domains [6]. But also the noisy phase can be improved. According to [7], a phase enhancement achieves a better result by about ΔMOS-LQO $= +0.2$ (*Mean Opinion Score, Listening Quality Objective* due to PESQ [8]) in noise scenarios with SNRs below 5 dB. In [9] a potential improvement of ΔMOS-LQO $= +0.1$ is assumed. However this work focuses on the envelope estimation which means no phase enhancement was done and obviously no approach for the excitation estimation, yet.

3 Signal Model

A time discrete microphone signal y_n can be described as addition of the desired speech signal s_n and noise d_n:

$$y_n = s_n + d_n. \tag{1}$$

3.1 Source-Filter Model

The source-filter model of speech production decomposes a speech signal into a filter and source component [10] wherein:

a) The vocal tract shapes the envelope, which is modeled as filter.
b) The human glottis generates the excitation signal, which is therefore called the source.

Speech production s_n is modeled as convolution of the envelope a_n with the excitation b_n:

$$s_n = a_n * b_n. \tag{2}$$

3.2 Analysis Filterbank

A *Short-Time Fourier Transformation* (STFT) with analysis window $h_{\text{ana},k}$ (Slepian-type) is applied to overlapped input segments according to:

$$\underline{Y}_{\mu,\eta} = \sum_{k=0}^{N_{\text{DFT}}-1} y_{\eta r-k} h_{\text{ana},k} e^{-j\Omega_\mu k}, \tag{3}$$

with $k = 0,1,...,N_{\text{DFT}}-1$, $N_{\text{DFT}} = 512$, frame shift $r = 128$, and $\Omega_\mu = \frac{2\pi}{N_{\text{DFT}}}\mu$ [11]. Moreover, $\mu = 0,1,...,N_{\text{DFT}}-1$ represents the subband indices and the frame index is denoted by η. The signals have a sampling frequency of $f_s = 16$ kHz. The short-term spectrum can be represented by $\underline{Y}_{\mu,\eta} = Y_{\mu,\eta} \cdot e^{j\varphi_{\mu,\eta}}$, wherein $Y_{\mu,\eta}$ is the magnitude spectrum and $\varphi_{\mu,\eta}$ is the phase. Due to the symmetry of the short-term spectrum in Eq. 3 only the first $M = \frac{N_{\text{DFT}}}{2} + 1$ subbands are used for feature extraction.

3.3 Spectral Feature Extraction

The envelope $A_{\mu,\eta}$ in spectral domain is extracted by means of a first order IIR-Filter applied on the magnitude spectrum $Y_{\mu,\eta}$. The implementation consists of a forward (recursively) and backward smoothing with a smoothing constant $\lambda = 0.8$:

$$Y'_{\mu,\eta} = \begin{cases} Y_{\mu,\eta}, & \text{if } \mu = 0, \\ \lambda Y'_{\mu-1,\eta} + (1-\lambda) Y_{\mu,\eta}, & \text{else.} \end{cases} \tag{4}$$

The backward smoothing is done analogues, yielding the spectral envelope:

$$A^{y}_{\mu,\eta} = Y''_{\mu,\eta}. \tag{5}$$

In addition, the excitation is extracted by element-wise division according to:

$$B^{y}_{\mu,\eta} = \frac{Y_{\mu,\eta}}{A^{y}_{\mu,\eta}}. \tag{6}$$

3.4 Feature Normalization

The logarithmized spectral envelope feature is $\alpha_{\mu,\eta} = 20\log_{10}(A^{y}_{\mu,\eta})$. The Eq. 7 shows how the normalization is applied [12]. It is done by mean subtraction followed by division of the standard deviation over each μ feature-bin, along every $\eta = 0,...,N-1$ time-frame of a whole data set:

$$\tilde{\alpha}_{\mu,\eta} = \frac{\alpha_{\mu,\eta} - \bar{\alpha}_{\mu}}{\sigma_{\alpha_{\mu}}}. \tag{7}$$

The normalization paradigm by choice is to normalize clean envelopes first and applying the obtained parameters on the noisy envelopes.

3.5 Reconstruction and Synthesis Filterbank

As a first step for the reconstruction of the enhanced signal the estimated spectral envelopes should be denormalized before multiplying it with the excitation. The normalization shown in section 3.4 is therefore done vice versa by multiplication of $\sigma_{\alpha_{\mu}}$ and adding $\bar{\alpha}_{\mu}$.

The denormalized and delogarithmized estimated envelopes are $\breve{A}_{\mu,\eta} = 10^{\frac{\breve{\alpha}_{\mu,\eta}}{20}}$. The reconstruction takes place by combining the excitation of the clean signal ($B^{s}_{\mu,\eta}$) and the estimated envelope again by element-wise multiplication over every subband for each frame:

$$\hat{Y}_{\mu,\eta} = B^{s}_{\mu,\eta}\,\breve{A}^{y}_{\mu,\eta}, \tag{8}$$

Signal synthesis is realized in a straightforward manner by first computing the inverse DFT of the reconstructed spectrum:

$$\hat{y}_{\kappa,\eta} = \frac{1}{N_{\mathrm{DFT}}} \sum_{\mu=0}^{N_{\mathrm{DFT}}-1} \underline{\hat{Y}}_{\mu,\eta}\, e^{j\Omega_{\mu}\kappa}, \tag{9}$$

and overlapping the time-domain signals $\hat{y}_{\kappa,\eta}$ after weighting with a synthesis-window (Slepian-type), wherein $\kappa = 0,1,2,...,N_{\mathrm{DFT}-1}$.

4 Deep-RNN

4.1 Architecture

Figure 1 shows the structure of the LSTM-based Deep-RNN architecture. The input of the Deep-RNN has a dimension of four spectral envelope frames with 257 features each. That means that four frames at a time get processed due to get one enhanced feature frame on the output, which is done by using a dense layer.

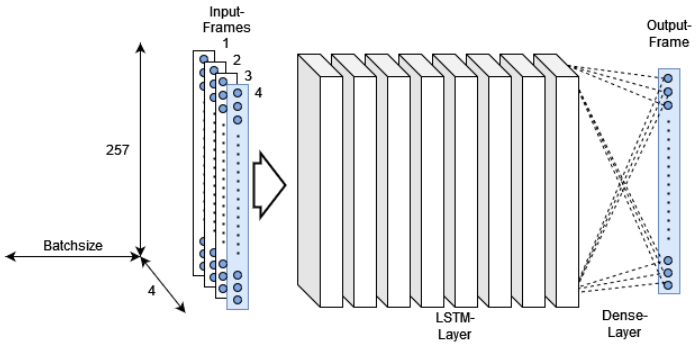

Figure 1 – Structure of the LSTM-based Deep-RNN architecture.

4.2 Data set

Noisy speech data for training and testing purposes is produced based on the 56 clean speakers of *Edinburgh Corpus* [13] by adding stationary car interior noise for several *Signal-to-Noise Ratios* (from -5 dB to 15 dB in 5 dB steps). In the same manner a 13 speaker test data set, with ≈ 0.5 h audio material was generated. It is detached from the training and validation sets, because there are no speakers of the training or validation sets included.

4.3 Training

The proposed Deep-RNN network is trained with the standard back-propagation [14], which applies the *Mini-Batch Gradient Descent* on a number of feature-frames N_{mb}. The *Mean Squared Error Loss* (MSE) serves as the objective function. The optimizer used is Adam with the learning rate 0.0002. The MSE is computed for log-normalized clean envelopes $\tilde{\alpha}^{s}_{\mu,\eta}$ and log-normalized estimated envelopes $\breve{\alpha}^{y}_{\mu,\eta}$, as follows:

$$L_{\mathrm{MSE}} = \frac{1}{N_{\mathrm{mb}} M} \sum_{\eta=0}^{N_{\mathrm{mb}}-1} \sum_{\mu=0}^{M-1} \left(\tilde{\alpha}^{s}_{\mu,\eta} - \breve{\alpha}^{y}_{\mu,\eta} \right)^2 . \tag{10}$$

During training the loss gets double checked by validation. The model parameters get saved, if training and validation loss decreased, otherwise the learning rate r_{learn} is lowered by 0.6, which is called *Performance Scheduling*. If validation differs significantly to training loss, the training runs into over- or underfitting. Training stops at this point to save computation time (*Early Stopping*). The ℓ_2*-Regularization* and *Dropout* were employed in order to reduce overfitting [12, 15]. Moreover a randomized audio-file order for each epoch was applied.

5 Speech Enhancement System

Figure 2 shows the concept of the idealized speech enhancement system. A noisy time signal y_n is transformed to spectral domain by *Analysis Filterbank* from section 3.2. The noisy phase is saved and used later in the back-transformation. The short-time spectrum $Y_{\mu,\eta}$ is smoothed by a IIR-Filter from section 3.3 to separate the envelope from excitation. Normalization is done by Eq. 7 on the logarithmized envelopes. The trained Deep-RNN provides the enhanced log-normalized envelopes $\breve{\alpha}^{y}_{\mu,\eta}$. Since in this study the objective is to determine the potential of the SE-System, there is no method used for excitation estimation. Instead the clean excitation $B^{s}_{\mu,\eta}$ and denormalized enhanced envelope $\breve{A}^{y}_{\mu,\eta}$ is used for reconstruction from section 3.5 to get an enhanced spectrum $\hat{Y}_{\mu,\eta}$. After back-transformation from section 3.5, the post-filter (Wiener Filter) is applied to deliver the further enhanced speech signal $\hat{s}_n$.

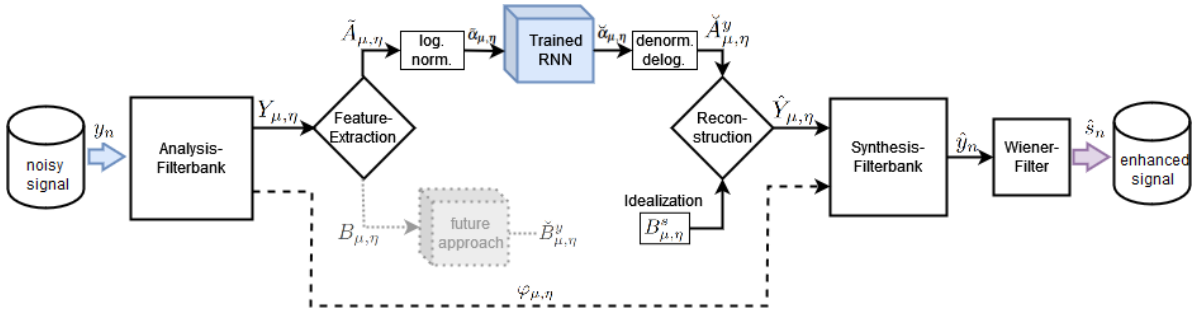

Figure 2 – Concept of the SE-System with idealized reconstruction.

6 Evaluation Methods

To evaluate and compare the performance of the SE-System in terms of speech quality and intelligibility, STOI and POLQA are utilized. STOI has a range between 0...100%. POLQA has a range between 0...4.75 (super-wideband 50-14000 Hz). Due to constraints of POLQA regarding file duration, only the valid share of the test data can be used. POLQA allows measurements with very high background noise and is therefore preferred over PESQ. To evaluate the quality on estimated feature level, the *Log Spectral Distance* combined with a *Speech Activity Detection* is used.

Speech Activity Detection: In order to apply measures separately, a *Speech Activity Detection* for binary frame masking is realized to distinguish speech-frames and noise-frames by $\overline{\mathrm{SNR}}_\eta$ of the current frame:

$$b_\eta = \begin{cases} 1 & \text{for } \overline{\mathrm{SNR}}_\eta > \vartheta, \\ 0 & \text{else,} \end{cases} \tag{11}$$

whereas for the noise power estimation the *Improved Minima Controlled Recursive Averaging Algorithm (IMCRA)* is used [16]. ϑ is set to 0.4. Speech-frames are selected by $\beta_\eta^{c=\text{speech}} = b_\eta$ and noise-frames are selected by $\beta_\eta^{c=\text{noise}} = 1 - b_\eta$. N_f is the number of all time-frames of same SNR, separated for each case c by $N_\mathrm{f}^c = \sum_{\eta=0}^{N_\mathrm{f}-1} \beta_\eta^c$.

Log Spectral Distance (LSD): The LSD [3] between a set of clean and estimated envelopes $A_{\mu,\eta}^s$ and $\hat{A}_{\mu,\eta}^y$ in spectral domain are calculated for the cases c over N_f as follows:

$$LSD^c = \frac{1}{N_\mathrm{f}^c} \sum_{\eta=0}^{N_\mathrm{f}-1} \beta_\eta^c \sqrt{\frac{1}{M} \sum_{\mu=0}^{M-1} \left[10 \log_{10} \left(\frac{A_{\mu,\eta}^s}{\hat{A}_{\mu,\eta}^y} \right) \right]^2}. \tag{12}$$

7 Results and Discussion

Architectural parameter adjustment and training parameter tuning were done for the used data set. Best found parameters for the architecture and training are in Table 1. The concatenation of eight LSTM-layers showed peak performance. The envelope estimation by the Deep-RNN shows high improvements at feature-level during speech with low SNRs as shown in Table 2. The STOI and MOS-LQ due to POLQA are in Table 3.

Behavior of the SE-System on different speakers: The Deep-RNN reacts differently to different speakers, which could yield the system suspicious being biased. This assumption is made by the experience by successive increasing the number of speakers during test and training:

Table 1 – Best found parameters for the used data set.

Parameter	Value
n_{Layer}	8
$n_{neurons}$	449
N_{mb}	512
$\ell 2$	0.002
dropout	0.6
r_{learn}	$2e^{-4}$
Optimizer	Adam
Loss	MSE

Table 2 – Average improvements by the Deep-RNN on the envelope feature by SNR.

	ΔLSD^c [dB]	SNR [dB]
speech noise	+6.01 +17.50	0
speech noise	+8.50 +20.00	-5

Table 3 – Average measures.

Scheme	$STOI$	$MOS-LQ$
1) Noisy data	0.747	2.04
2) Wiener Filter	0.747	2.02
3) Reconstruction	0.870	2.88
4) SE-System	0.862	2.85

the results tendency to get better or worse fluctuated. However, the cause could be due to the speaker themself. On that account a file-randomization was used, which showed a slight overall improvement, but wasn't separately examined on the bias hypothesis. In general, a test for diversity should be derived in the future, which may require the use of international speech data base including pitch sensitive tonal languages for a true speaker generalization, similar to [17].

8 Conclusion

The training was improved by file-randomization on a batch, every epoch. Also the frameworks for data generation, training and evaluation were improved to handle the amount of data. The architecture was successfully tuned to fit the requirements for a more intrusive training and testing scenario to approach the speaker generalization. The quality at feature level shows nearly as good results as in the single speaker scenario before. Moreover, the full SE-System with post-filter was tested in an idealized approach which assume the excitation signal to be perfectly estimated. In that way the potential of the SE-System with envelope estimation due to the proposed Deep-RNN is shown. Especially with bigger data sets and several noise scenarios, architectural modification will be necessary, because the number of eight LSTM layers seem to exploit the performance for the used data sets. A more sophisticated architecture, for example using bidirectional LSTMs, could improve the results. Furthermore, there is a need for a suitable approach on excitation estimation to complete the SE-System in a first stage. Due to insights by successive increasing the number of speakers during parameter tuning, there is a need for sophisticated test on speaker diversity found.

References

[1] KRINI, M. and G. SCHMIDT: *Speech and Audio Processing in Adverse Environments*, chap. Model-based Speech Enhancement, pp. 89–134. E. Hänsler, G. Schmidt (eds.), The address of the publisher, 1 edn., 2008.

[2] CIBA, S. and M. KRINI: *Spectral envelope estimation using deep recurrent neural networks for speech reconstruction. Proceedings of 47th German Annual Conference on Acoustics (DAGA)*, 2021.

[3] GRAY, A. H. and J. D. MARKEL: *Distance measures for speech processing. IEEE transactions on acoustics, speech, and signal processing (ASSP), vol. 24, no. 5 , October*, 1976. `https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=1162849`.

[4] C. H. TAAL, R. H., R. C. HENDRIKS and J. JENSEN: *An algorithm for intelligibility*

prediction of time-frequency weighted noisy speech. IEEE Transactions on Audio, Speech, and Language Processing, vol. 19, no. 7, pp. 2125-2136, september, 2011.

[5] OPTICOM: *Perceptual objective listening quality analysis*. Technical White Paper, 2011.

[6] S. ELSHAMY, W. T., N. MADHU and T. FINGSCHEIDT: *DNN-supported speech enhancement with cepstral estimation of both excitation and envelope. IEEE/ACM Transactions on Audio, Speech and Language Processing, vol. 26, no. 12, pp. 2460 - 2474*, 2018.

[7] MOWLAEE, P. and R. SAEIDI: *On phase importance in parameter estimation in single-channel speech enhancement. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2013.

[8] (ITU), I. T. U.: *Rec. p.862: Perceptual evaluation of speech quality(pesq): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speechcodecs*. 2001.

[9] GERKMANN, T., M. KRAWCZYK, and R. REHR: *Phase estimation in speech enhancement - unimportant, important, or impossible? IEEE Convention of electrical and electronics engineers in Israel*, 2012.

[10] VARY, P., U. HEUTE, and W. HESS: *Digitale Sprachsignalverarbeitung*. N. Fliege and M. Bossert, Teubner Stuttgart, 1 edn., 1998.

[11] HÄNSLER, E. and G. SCHMIDT: *Acoustic Echo and Noise Control - A Practical Approach*. E. Hänsler, John Wiley & Sons, Hoboken, NJ, USA, 1 edn., 2004.

[12] GÉRON, A.: *Hands-On Machine Learning with Scikit-learn and Tensorflow*. O'Reilly Media, O'Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472, 2 edn., 2019.

[13] VALENTINI-BOTINHAO, C.: *Noisy speech database for training speech enhancement algorithms and tts models, 2016 [sound]*. University of Edinburgh. School of Informatics. Centre for Speech Technology Research (CSTR), 2016. URL: `http://dx.doi.org/10.7488/ds/2117`.

[14] RUMELHART, D. E., G. E. HINTON, and R. J. WILLIAMS: *Learning representations by back-propagating errors*. *Nature*, 323, pp. 533–536, 1986.

[15] GAL, Y. and Z. GHAHRAMANI: *A theoretically grounded application of dropout in recurrent neural network. ArXiv e-prints*, 2015. ArXiv:1512.05287.

[16] COHEN, I.: *Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging*. *IEEE, Transactions on Speech and Audio Processing*, 11(5), pp. 466 – 475, 2003.

[17] KOUNOVSKY, T. and J. MALEK: *Single channel speech enhancement using convolutional neural network. IEEE, International Workshop of Electronics, Control, Measurement, Signals and their Application to Mechatronics (ECMSM)*, (5), pp. 1–5, 2017. DOI: 10.1109/ECMSM.2017.7945915.