DNN-based Speech Enhancement with Harmonics Regeneration

Patrick Preißner[a], Jonas Sautter[a], Friedrich Faubel[a], Markus Buck[a], Wolfgang Minker[b]
[a] Cerence Inc., Email: patrick.preissner@cerence.com
[b] Institute of Communications Engineering, University Ulm

Abstract
In a hands-free telephony setting for automotive environments, driving noise distorts both the phase and amplitude of acquired speech signals. The aim of noise reduction approaches is to restore clean speech before sending it over the telephone channel. State-of-the-art approaches use DNN-based filters that estimate real or complex–valued filter coefficients in the frequency domain. While real-valued filters can only correct the spectral magnitude, complex-valued filters can also correct the phase. But the latter comes at a significantly higher computational cost. In this work, we take an alternative approach that re-synthesizes the harmonic structure of speech by using DNN-based pitch trackers and voiced/unvoiced detectors. The re-synthesis concentrates on the frequency band between 0 and 1000 Hz as the human hearing system is most sensitive to phase errors in this range. Higher frequency bands of the speech signal are obtained with a real-valued filter whose coefficients are estimated with a DNN. The perceived speech quality of the processed speech is evaluated in subjective listening tests.

Introduction
Single channel speech enhancement is a widely studied topic. It had its beginnings in the 1970th with the seminal works by Callahan [1] and Boll [2] that led to a decades lasting era of statistical signal processing approaches. Follow-up work investigated the use of a short-time spectral magnitude estimator [3], minimum statistics based noise estimation [4] and a-priori signal to noise estimation [4]. Later approaches made use of an explicit model of the clean speech distribution [5, 6], non-negative matrix factorization (NMF) [7] and countless other ideas.

In the last decade, however, the era of statistical signal processing has increasingly been replaced by the era of deep learning. In this new era, ideal ratio masks (IRM) [8] or the Wiener filter supression rule [9] are directly estimated from the noisy input signal, by using a deep neural network (DNN). Follow-up work extended this to complex ratio masks [10, 11] that do not only correct the spectral magnitude but also the phase information.

The argument in [10] that the phase matters is supported by a much more detailed analysis in [12] which strongly suggests that phase errors significantly reduce the perceived quality of speech in MOS evaluations. In particular, it is shown that the perception of the lowest harmonics is most strongly affected by phase errors. The largest perceived difference occurs for low fundamental frequencies around 50 Hz and it decays with increasing frequency until it is barely noticeable around 800 Hz.

Motivated by these findings, we investigate the combination of DNN-based noise suppression with harmonic reconstruction. The idea is to avoid complex-valued networks that require twice the number of nodes and four times the computational expense. Hence, speech is reconstructed with sinusoidal synthesis. In contrast to full-band speech synthesis [13], only the lower frequencies are reconstructed where strong degradations due to phase errors are expected.

Overview:

The remainder of the paper is organized as follows: The upcoming section briefly reviews DNN-based Wiener filtering plus the architecture used in this work. This is followed by an in-depth discusssion of the harmonic regeneration approach, which is finally evaluated in the experimental section. The paper is wrapped up with the conclusion.

DNN-based Wiener filter
State-of-the-art noise suppression approaches use DNN-based spectral weighting to enhance the subjective quality of speech signals. This is achieved by estimating real-valued weights $\hat{H}(l,k) \in [0,1]$ and then multiplying these weights to the complex-valued spectral coefficients $Y(l,k)$ of the input signal $y(n)$:

$$\hat{X}(l,k) = Y(l,k) \cdot \hat{H}(l,k). \tag{1}$$

In this equation, l and k denote time and frequency indices, respectively. Y denotes the STFT of the noisy microphone signal y. And $\hat{X}$ is the estimated clean speech spectrum. A weight of 1 means that the current spectral component is speech and should be preserved. A weight of 0 means the component is noise and should be removed.

Similiar to classical noise reduction methods, different suppression rules like Wiener filter [1, 9] or spectral masks [14] may be used. However, the suppression rules are not applied directly but just used as a "target" that the neural network is supposed to learn. For this to work, clean speech and noise sepctra X and N need to be mixed synthetically, as shown in Figure 1. The result of the mixing step is noisy input spectrum Y from which the ideal spectral weights can be calculated because all X, N and Y are known. In case of the Wiener filter, which is used in this work, the ideal weights are:

$$H_{\text{opt}}(l,k) = 1 - \frac{|N(l,k)|^2}{|Y(l.k)|^2}. \tag{2}$$

These target weights are supposed to be learned in the DNN training stage, i.e. they are estimated at the output

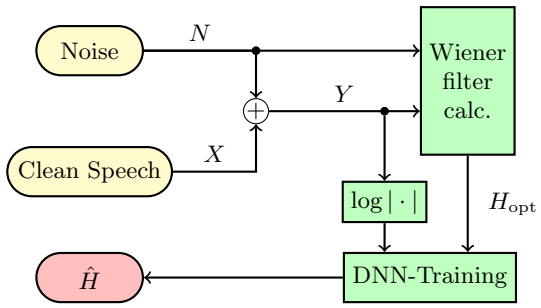

Figure 1: Training setup for DNN-based Wiener filter.

layer of the network based on the log spectral magnitude values $\log|Y|$ of Y that are presented at the network input. Like in a typical regression task, the objective of the training is to minimize the loss between network output $\hat{H}$ and target H_{opt} in a minimum mean squared error (MMSE) sense:

$$\mathcal{L}_{\text{MSE}}(\hat{H}, H_{\text{opt}}) = ||\hat{H} - H_{\text{opt}}||_2^2. \tag{3}$$

Figure 2 shows the network architecture used in this work. At its core, the network consists of two recurrent (RNN) layers that model the temporal behavior of speech and noise over time. These are sandwiched by feed-forward fully-connected (FC) layers (also called Dense layers) towards the input and output layers. Similar as originally proposed in [15], the RNNs are implemented as Gated Recurrent Units (GRU). Additionally, residual connections [16, 17] are used to bypass the GRU layers and then add the GRU input to its output. This was found to significantly improve the network performance in preliminary experiments. As we use an FFT of size 512, both the input and output layers of the network have 257 nodes. All the hidden layers use 256 units.

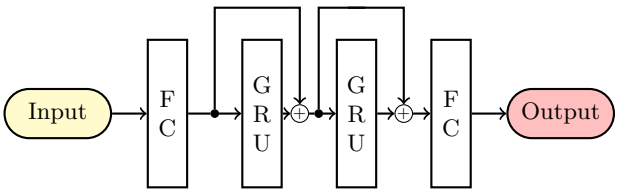

Figure 2: Architecture of the noise suppression network.

Harmonics Regeneration

Since the denoising of speech signals with a Wiener filter is a magnitude spectrum based method [1], the phase spectrum of the signal is not changed. Previous work [12] has shown that the perceived speech quality degrades not only with distortions to the magnitude-spectrum but also with distortions to the phase-spectrum in lower frequency components (1000 Hz and below).

Figure 3 and figure 4 demonstrate the effect of low frequency driving noise on the phase continuity at the fundamental frequency of human speech. The introduced jitter in the phase of the noisy speech signal is especially visible in the highlighted area between 0.4 s and 0.7 s.

For the complex spectrum $Y(l,k)$, the delta-phase $\Delta\phi(l)$ with respect to the timestep l is then calculated according to

$$\Delta\phi(l) = \arg\left(Y\left(l, k_{f_0}(l)\right) \cdot Y^*\left(l+1, k_{f_0}(l+1)\right)\right) \tag{4}$$

with the complex argument $\arg(z)$ for any number $z \in \mathbb{C}$. The spectral component $k_{f_0}(l)$ is the closest frequency bin to the fundamental frequency $f_0(l)$ at frame l and is given by

$$k_{f_0}(l) = \left[\frac{f_0(l) \cdot N}{F_s}\right] \tag{5}$$

where N is the number of data points used for the fast Fourier transform and F_s is the sampling frequency.

The proposed approach to correct the delta-phase through harmonics regeneration is realized with a DNN-based pitch tracking and voiced/unvoiced decision.

Pitch Tracker

The pitch tracker network follows the same architecture as the DNN based Wiener filter, except that the hidden layers have size 200. The output layer consists of a single node.

Since the pitch is a continous value $f_0 \geq 0$, the learning task is a regression problem. To generate the target f_0, RAPT [18] was applied on the clean speech signals.

Following RAPT, we define the pitch to be 0 for unvoiced and non speech parts in the input data. In order to not introduce a bias in the training stage, the pitch tracking network is only trained on voiced frames. It is left to a separate voiced / unvoiced decision to set the final pitch estimate to 0. Contrary to the Wiener filter network from the previous section, the pitch tracking network uses the mean absolute error (MAE) as a loss function:

$$\mathcal{L}_{\text{ThrMAE}}(\tilde{f}_0, f_0) = \begin{cases} 0 & \text{if } f_0 = 0 \\ |\tilde{f}_0 - f_0| & \text{if } f_0 > 0 \end{cases} \tag{6}$$

Voiced/Unvoiced Decision

As stated in the last paragraph, the pitch estimation network is trained on voiced frames of the speech signal. Consequently the pitch tracker output can not be expected to equal 0 for unvoiced or non-speech parts. Hence, we train an additional voiced/unvoiced decision (VUD) network that classifies each frame to be either voiced or unvoiced.

Just like the pitch tracker, the VUD uses the same architecture as in figure 2. All hidden layers are 128 nodes wide. The single output node of the network is denoted by p_{voiced} and lies within the range $[0, 1]$. As for most binary classification tasks, the cross-entropy is used as the loss function for the training. For each frame the input of the network is its logarithmic magnitude spectrum. We then define the frame l to be voiced if $p_{\text{voiced}} > 0.5$ for the given input feature and unvoiced otherwise. This can be described by a binary mask

$$m(l) = \begin{cases} 1 & \text{if } p_{\text{voiced}}(l) > 0.5 \\ 0 & \text{if } p_{\text{voiced}}(l) \leq 0.5 \end{cases} \tag{7}$$

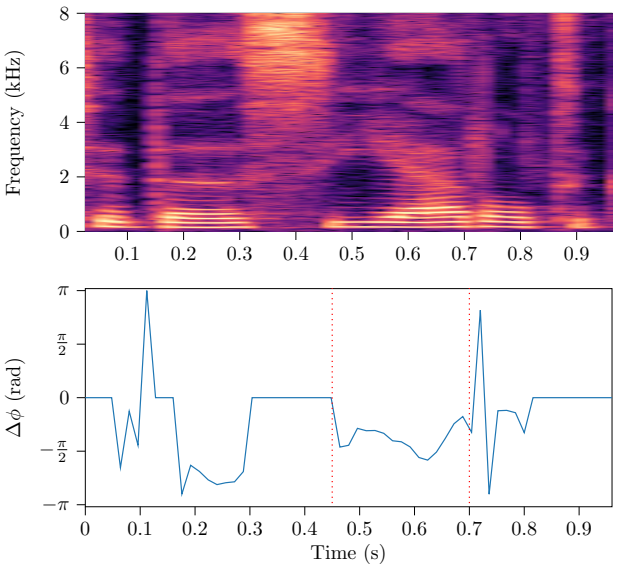

Figure 3: Magnitude spectrum of clean speech and corresponding delta-phase at f_0.

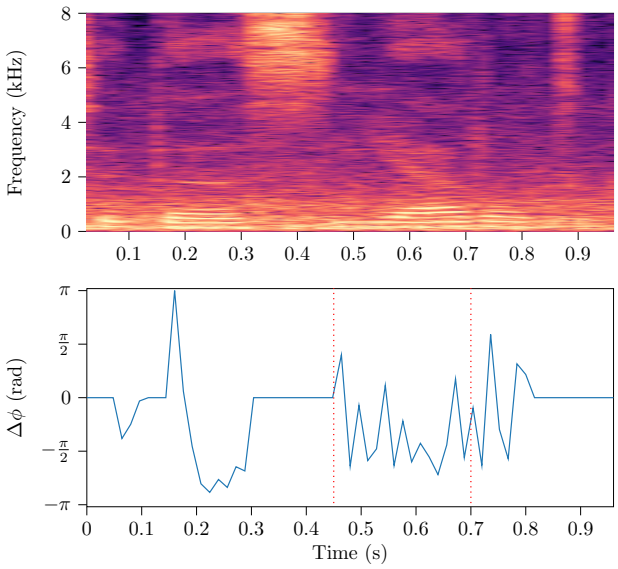

Figure 4: Magnitude spectrum of noisy speech with additive driving noise and corresponding delta-phase at f_0.

The resulting estimated pitch contour for a speech signal is then given by

$$\hat{f}_0(l) = \tilde{f}_0(l) \cdot m(l). \tag{8}$$

Sinusoidal Synthesis

Using the estimated fundamental frequency $\hat{f}_0(l)$ we can now resynthesize the harmonic structure in the lower frequency bands of voiced speech. For this, we assume $\hat{f}_0(l)$ is the pitch at the center of the frame l. To model a smooth transition between adjacent frames, the linear interpolation $\hat{f}_{0,s}(n)$ of the pitch is calculated for each time-domain sample index n. Subsequently, the excitation consisting of the fundamental frequency and its harmonics is re-synthesized as follows:

$$g_{\mathrm{Exc}}(n) = \sum_{m=1}^{M} \sin \phi_m(n), \tag{9}$$

where

$$\phi_m(n) = \phi_m(n-1) + \frac{2 \cdot \pi \cdot m \cdot \hat{f}_{0,s}(n)}{F_s}. \tag{10}$$

M is chosen to be sufficiently large for the generated harmonics to cover the whole frequency range up to the cutoff frequency $f_c = 500\,\mathrm{Hz}$.

To model the amplitude spectrum of the speech, we assume at max a linear decay of the energy from f_c down to 0 Hz. Consequently, the spectral envelope for the generated harmonics is approximated as follows:

$$G_{\mathrm{Env}}(l,k) = \max\left(\frac{k \cdot \hat{X}_{\mathrm{Env}}(l,k_{f_c})}{k_{f_c}}, \hat{X}_{\mathrm{Env}}(l,k)\right), \tag{11}$$

where k_{f_c} denotes the frequency bin corresponding to the cutoff frequency f_c. Multiplying by the excitation G_{Exc} in the frequency domain gives the re-synthesized speech spectrum:

$$G(l,k) = G_{\mathrm{Exc}}(l,k) \cdot G_{\mathrm{Env}}(l,k). \tag{12}$$

The enhanced signal is finally obtained by cross-fading G into the spectrum $\hat{X}$ of the filtered speech signal, starting at the cut-off frequency f_c.

Experiment

A listening test was conducted to compare the subjective speech quality between signals processed only with the DNN-based Wiener filter as well as with additional harmonics regeneration.

All three networks - for the Wiener filter, pitch tracker and VUD - were trained with a total of 40 h of speech from the SPEECON corpus [19]. The data was randomly selected from the "office" condition of four different languages: Dutch-NL, English-US, French-FR and German-DE. Automotive driving noise and wind bursts were used as a noise source that was added to the clean speech data. To prevent overfitting druing training, we used L_2-Regularization with $\lambda = 10^{-4}$ and dropouts [20] for all hidden layers with a rate of $p = 0.1$. Rectified Linear Unit (ReLU) were used as an activation function for fully-connected layers. GRU layers were using tanh for forward activations and the hard sigmoid function for recurrent activations.

The evaluation dataset for the listening test was randomly chosen from the SpeeCon corpus. It consists of 12 utterances from male speakers and 12 utterances from female speakers that were not part of the training data set. These speech files were mixed with driving noise and wind bursts at an SNR of 5 dB and then processed with the DNN-based Wiener filter with and without harmonics generation. The difference between the processed files was rated in a listening test with 16 participants. Figure 5 shows the results in terms of the comparison mean opinion score (CMOS).

For male speakers, most participants slightly prefered the processed signals with additional harmonics regeneration. For this case the average CMOS was 0.21 and

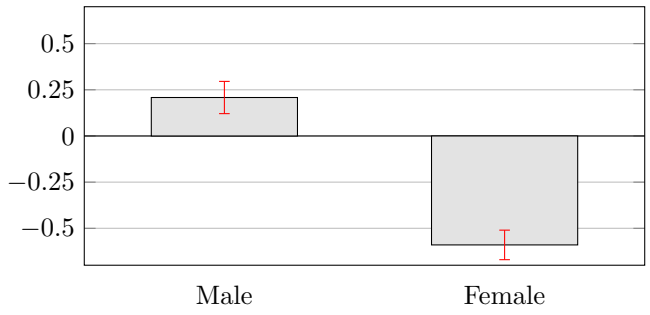

Figure 5: Comparison mean opinion score for files processsed with and without harmonics regeneration. Results are shown individually for male and female speakers, with standard error bars. Note that the scale of the y-axis is in the range $[-0.5, 0.5]$ instead of the full CMOS range $[-3, 3]$.

the standard error was 0.12. For female speakers, the listeners preferred speech signals without the additional processing step. Here, the average CMOS was -0.59 and the standard error was 0.085.

Conclusion

In this paper, we evaluated a combined approach for noise suppression and harmonics regeneration. In this approach, a DNN-based Wiener filter is used to denoise the signal and estimate the spectral envelope. Low frequency speech components are reconstructed by first estimating the pitch contour with a DNN-based pitch tracker and then re-synthesizing the harmonic structure below 500 Hz with sinusoidal synthesis. Subjective listening tests confirmed that this approach enhances the subjective speech quality for utterances spoken by male speakers. Unfortunately, it also reduces the subjective quality for female speakers. Further work might investigate the reason for the discrepancy.

References

[1] Callahan, M.: Acoustic signal processing based on the short-time spectrum, Acoustical Society of America Journal (1977)

[2] Boll, S.: Suppression of acoustic noise in speech using spectral subtraction, IEEE Transactions on Acoustics, Speech, and Signal Processing (1979), 113-120

[3] Ephraim, Y. & Malah, D.: Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator, IEEE Transactions on Acoustics, Speech, and Signal Processing (1984)

[4] Martin, R.: Noise power spectral density estimation based on optimal smoothing and minimum statistics, IEEE Transactions on Speech and Audio Processing (2001), 504-512

[5] Lotter, T. & Vary P.: Noise Reduction by Joint Maximum A Posteriori Spectral Amplitude and Phase Estimation with Super-Gaussian Speech Modelling (2004)

[6] Hadir et al.: A Model-Based Spectral Envelope Wiener Filter for Perceptually Motivated Speech Enhancement, Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH (2011), 213-216

[7] Mohammadiha et al.: A new approach for speech enhancement based on a constrained Nonnegative Matrix Factorization, International Symposium on Intelligent Signal Processing and Communications Systems (ISPACS) (2011), 1-5

[8] Narayanan, A. & Wang, D.: Ideal ratio mask estimation using deep neural networks for robust speech recognition, EEE International Conference on Acoustics, Speech and Signal Processing (2013), 7092-7096

[9] Mirsamadi, S. & Tashev, I.: A Causal Speech Enhancement Approach Combining Data-driven Learning and Suppression Rule Estimation, Proc. InterSpeech 2016 (2016)

[10] Williamson et al.: Complex Ratio Masking for Monaural Speech Separation, IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, 2016

[11] Xia2017, S: Using Optimal Ratio Mask as Training Target for Supervised Speech Separation (2017)

[12] Laitinen et al.: Sensitivity of Human Hearing to Changes in Phase Spectrum. AES: Journal of the Audio Engineering Society (2013), 860-877

[13] Al-Radhi et al.: RNN-based speech synthesis using a continuous sinusoidal model, CoRR (2019)

[14] Holmes, J. & Sedgwick, N.: Noise compensation for speech recognition using probabilistic models, Proceedings of the 1986 IEEE International Conference on Acoustics, Speech and Language Processing (ICASSP '86), vol. 11 (1986), 741-744

[15] Cho et al.: Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation (2014)

[16] Toderici et al.: Full Resolution Image Compression with Recurrent Neural Networks, 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017), 5435-5443

[17] He et al.: Deep Residual Learning for Image Recognition, IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016), 770-778

[18] Talkin, D.: A Robust Algorithm for Pitch Tracking (RAPT). Speech Coding and Synthesis (2015), 497-518

[19] Iskra et al.: SPEECON - Speech Databases for Consumer Devices: Database Specification and Validation, Proceedings of LREC (2002)

[20] Srivastava et al.: Dropout: A Simple Way to Prevent Neural Networks from Overfitting, Journal of Machine Learning Research (2014), 1929-1958