INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

On the influence of modifying magnitude and phase spectrum to enhance noisy speech signals

Hans-Günter Hirsch[1], *Michael Gref*[1,2]

[1]Institute for Pattern Recognition, Niederrhein University of Applied Sciences, Krefeld, Germany
[2]Fraunhofer Institute IAIS, St. Augustin, Germany

hans-guenter.hirsch@hs-niederrhein.de, michael.gref@hs-niederrhein.de

Abstract

Neural networks have proven their ability to be usefully applied as component of a speech enhancement system. This is based on the known feature of neural nets to map regions inside a feature space to other regions. It can be taken to map noisy magnitude spectra to clean spectra. This way the net can be used to substitute an adaptive filtering in the spectral domain. We set up such a system and compared its performance against a known adaptive filtering approach in terms of speech quality and in terms of recognition rate. It is a still not fully answered question how far the speech quality can be enhanced by modifying not only the magnitude but also the spectral phase and how this phase modification could be realized. Before trying to use a neural network for a possible modification of the phase spectrum we ran a set of oracle experiments to find out how far the quality can be improved by modifying the magnitude and/or the phase spectrum in voiced segments. It turns out that the simultaneous modification of magnitude and phase spectrum has the potential for a considerable improvement of the speech quality in comparison to modifying the magnitude or the phase only.

Index Terms: speech enhancement, deep neural networks, modification of magnitude and phase spectrum, pitch synchronous analysis

1. Introduction

Deep neural networks (DNN) are already used as a state-of-the-art component in automatic speech recognition (ASR) for some years. In the past few years they also led to remarkable improvements in the field of speech enhancement. Very promising approaches that significantly outperform classical adaptive filtering are presented, e.g. in [1]. Neural networks are usually trained to map the logarithmic magnitude spectra of noisy signals to the corresponding magnitude spectra of the clean signal where the spectral analysis is done by means of a Short-Time Fourier Transform (STFT). The optimization of this spectral estimation approach includes for example the choice of an appropriate network topology and appropriate training parameters, the expansion of the training data to consider further noise scenarios, the usage of further features as input to the net and the application of a multi-objective-learning to jointly estimate secondary target parameters [1], [2]. Most often, the phase information is not taken into account. An overview about phase-aware signal processing is given in [3]. To obtain the enhanced time domain signal, the STFT-phase of the noisy signal is used along with the STFT magnitude spectra estimated by the DNN.

This paper aims to analyze the influence of modifying the magnitude spectra and the phase spectra for DNN-based speech enhancement methods. The potential for improvement in different applications by modifying the STFT phase spectra has been shown in different studies, e.g. [4], [5]. A method is presented in [6] for the reconstruction of the STFT-phase in voiced segments only. The method intends to provide a consistent phase in consecutive speech frames of the STFT analysis scheme based on a harmonic model of speech production. An approach of combining the estimation of the magnitude spectra by means of a DNN and a STFT-phase reconstruction has been investigated in [7]. This phase-reconstruction method is mainly based on the well-known Griffin-Lim-Algorithm [8]. Therefore, its efficiency is mainly determined by the quality of the magnitude spectrum estimation. Overall, the influence of a phase modification - especially its potential in DNN-based speech enhancement methods - is still an open issue that is not completely clear.

In this paper we would like to contribute to the clarification of this issue. We present a DNN based method and its efficiency to enhance noisy speech by modifying the magnitude spectrum in the STFT domain. Its potential is evaluated by measuring the speech-quality with the ITU-T recommended algorithm for the *perceptual evaluation of speech quality* (PESQ) [9] where the PESQ-MOS value is used as quantitative measure [10]. On the other hand, we evaluate the noise reduction capabilities by running a set of recognition experiments on the Aurora-4 corpus [11]. As basis for the DNN-HMM-based ASR we use the Kaldi framework [12]. We investigate the capability for enhancing speech with a set of orcale experiments where we substitute the noisy spectral magnitude or/and the spectral phase by the corresponding components of the clean signal.

2. DNN-based enhancement using magnitude modification in the STFT-domain

Similar to [1] we train deep neural networks to map the logarithmic short-term magnitude spectra of noisy signals to the corresponding spectra of the clean signal. The STFT of a signal $x[n]$ is thereby defined as the Discrete Fourier Transform (DFT) of equally windowed, overlapping segments of the signal. For a window-function $w[n]$, a hop size $H \in \mathbb{N}$ and a DFT-length $N \in \mathbb{N}$ this can be defined as

$$(X_{k,m})_{k=0}^{N-1} := \mathcal{DFT}\left((x[n+m\cdot H]\cdot w[n])_{n=0}^{N-1}\right). \quad (1)$$

The DFT frequency index is $k \in \mathbb{Z}$. $m \in \mathbb{Z}$ is the time index referring to an individual windowed segment. At a sample frequency $f_s = 16$ kHz we use a Hamming window with 400 samples length, a DFT-length $N = 512$ and a hop size $H = 160$. We use three successive log-magnitude spectra with $N/2+1 = 257$ components each as input to the neural net. As described as "noise aware training" in [13] we combine the three spectra with an spectral estimate of the background noise. We apply a method [14] to estimate the logarithmic noise spectrum under the assumption of a stationary noise signal in the background. This estimate is also needed to realize the classi-

http://dx.doi.org/10.21437/Interspeech.2017-1173

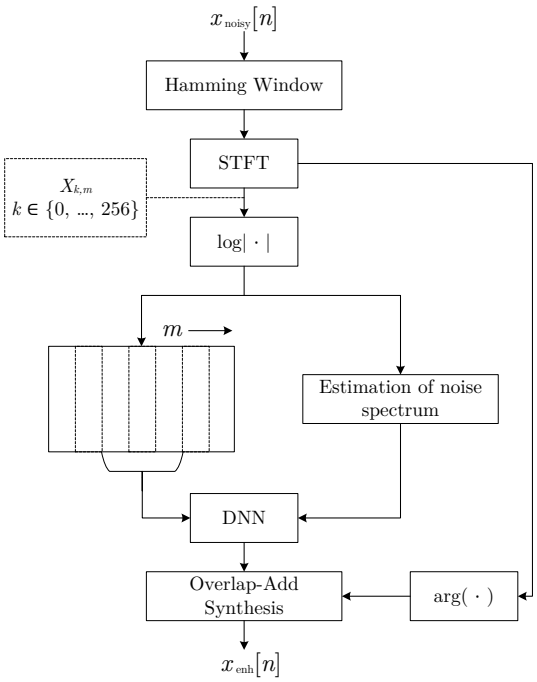

Figure 1: *Schematic structure of the DNN-based noise reduction method*

cal adaptive filtering approaches. For training the weights of the neural net we provide the logarithmic magnitude spectra of the corresponding segments of the clean signal as target values at the output of the net. We assume the parallel availability of a clean and a distorted version of the same signal as it is the case in the Aurora-4 database that we will describe in the following subsection. The schematic structure of the noise reduction method is illustrated in figure 1. For the reconstruction of the enhanced signal we use the phase of the noisy signal.

2.1. Usage of the Aurora-4 corpus

In this work we use signals of the Aurora-4 corpus [11], [15] that contain distorted versions of clean signals from the WSJ0-Corpus (Wall Street Journal) [16], [17]. The noisy signals have been artificially created by adding noise signals of six different scenarios (*airport, babble, car, restaurant, street* and *train*) to clean signals. The Aurora-4 framework can be used to run experiments and achieve comparable recognition results in the field of robust recognition. A 14 h training set is available for training a recognition system in multi-condition mode that includes noisy data for the six noise scenarios with random SNRs between 10 and 20 dB. This 14 hour set with a total of 5.4 million short-term spectra is used to train the neural networks. For evaluation Aurora-4 includes a test set with 330 speech signals from speakers not used in the training set. Distorted versions of the 330 signals are available for the six noise conditions with a randomly selected SNR between 5 and 15 dB. To evaluate the robustness of the noise reduction method we created four further noisy versions as an additional test set for the scenarios *bus, ICE-train, lobby* and *local-train* using randomly applied SNRs in the same range as in the original test-set. Thus, we want to cover the case of noisy scenarios that have not been used while training the neural net. We refer to this as the *additional test-set* to distinguish it from the standard Aurora-4 test-set. In this paper we work with signals sampled at a frequency $f_s = 16$ kHz. For measuring speech quality we process the whole set of 330 signals in each noise condition and calculate the average PESQ value. For recognition, we apply the example script for the Aurora-4 corpus as provided within the Kaldi framework to achieve comparable and reproducible results.

2.2. Training of deep neural networks for noise reduction

The deep neural network applied as component of the noise reduction scheme is trained with two different frameworks. Thus, we want to compare both frameworks with respect to their suitability for this task of mapping noisy spectra to clean ones. First, we apply the so called NNET1 training procedure which is a tool within the Kaldi framework that is usually taken to train a neural network as component of a recognition system. Alternatively we apply the training tool of Keras using TensorFlow as the backend. To distinguish the aforementioned Kaldi DNN-based ASR and this noise reduction DNN we refer to the first one as *Kaldi DNN-HMM-ASR* and to the second one simply as the *Kaldi DNN*.

The objective function used for training is the mean squared error. The learning rate is adaptively halved if the validation loss is not decreasing anymore. The exact ruling for this varies due to the different implementations in Kaldi and Keras. At our implementation in Keras the learning rate usually decreases later than in the Kaldi NNET1-implementation. The initial learning rate value is 10^{-5} in both cases. In Kaldi a mini-batch Stochastic Gradient Descent (SGD) training is used. In Keras we used the RMSprop optimizer instead. The mini-batch size in both cases is 256. In general, all hidden layers of a network contain the same number of neurons. tanh is used as the activation function. The activation at the output layer is the identity function. While training we varied the number of hidden layers as well as the number of neurons per layer to find the neural network configuration with the lowest validation loss.

Training with Keras, the loss continuously decreases for increasing network size. We looked at a maximum number of 8 hidden layers and a maximum number of 2000 neurons per layer to limit the computational load. We would expect further small improvements for a still increasing network size. However, using Kaldi we find an optimum at a network size of 3 hidden layers with 600 nodes per layer. Further increasing or decreasing the size leads to a worse validation loss.

2.3. Results using the STFT-phase of the noisy signal for synthesis

Calculating the PESQ-MOS on the signals of the test sets after applying a classical adaptive filtering or the aforementioned DNN-based noise reduction we achieved the results listed in table 1. The shown numbers are the mean PESQ values over the six noise scenarios included in Aurora-4 respectively over the four additional noise conditions. The adaptive filtering has been developed as part of earlier work [18]. It contains a cepstral smoothing technique [19] to reduce the amount of musical tones. We achieve higher PESQ mean values with the DNN-based enhancement methods in comparison to the adaptive filtering approach. The use of the Keras DNN results in higher PESQ values than the Kaldi DNN probably due to its larger network size.

As expected, the improvement in speech quality by the DNN-based methods is lower for the noise scenarios of the additional test set as for the Aurora-4 scenarios that have been

Table 1: *Mean values of PESQ-MOS for noisy signals enhanced by different methods*

	Aurora-4	**Additional Set**
Noisy (no enhancement)	1.971	2.004
Adaptive Filtering	2.264	2.283
DNN (Kaldi)	2.513	2.359
DNN (Keras)	2.615	2.429

used for training the neural net. This indicates an insufficient generalization to noise types not seen in the training phase. In table 2 the results are analyzed a bit more in detail by listing the mean PESQ values for a few selected noise types. The key benefit of DNN-based noise reduction becomes obvious. The DNN-based methods are just slightly better than the adaptive filtering in case of stationary noise types like *car* or *ICE*. But for non stationary noise types like *babble* or *lobby* the improvement is significantly higher.

Table 2: *Selected PESQ-MOS mean values for noise types in the Aurora-4 test set (middle) and the additional test set (right)*

	clean	**babble**	**car**	**ICE**	**lobby**
Noisy	4.549	1.936	2.315	1.971	1.902
Adap. Fil.	4.445	2.191	2.760	2.269	2.162
DNN (Kaldi)	3.994	2.441	2.899	2.286	2.245
DNN (Keras)	4.103	2.539	3.031	2.370	2.307

The advantage of the DNN based approach over the adaptive filtering can also be observed in automatic speech recognition. The methods are applied to create noise reduced speech signals that are taken as input to a Kaldi DNN-HMM-ASR recognition system. In table 3 the word error rates are listed for the case of training the neural net inside the recognition system on clean data only. We observe already a considerable reduction of the error rates when applying the adaptive filtering. But these results are outperformed by the ones achieved with the DNN based enhancement. Again the Keras DNN leads to better results than the Kaldi DNN. Both perform better on the Aurora-4 test set than on the additional set.

Table 3: *Word-error-rates of a clean-trained Kaldi DNN-HMM-ASR for noisy signals enhanced by different methods*

	Aurora-4	**Additional Set**
Noisy (no preprocessing)	53.4 %	42.7 %
Adaptive Filtering	33.2 %	28.3 %
DNN (Kaldi)	12.0 %	19.3 %
DNN (Keras)	11.0 %	17.2 %

Usually a multi-condition training is applied for increasing the robustness. For the sake of completeness the word error rates are listed in table 4 for the DNN-HMM-ASR system trained in multi-condition mode. As expected, the error rates are higher for the enhanced signals than for the noisy signals due to the double application of the DNN as part of the enhancement method and as component of the recognition framework and due to the small artificial distortions that are introduced by the enhancement processing.

Table 4: *WER of a multi cond. trained Kaldi DNN-HMM-ASR*

	Aurora-4	**Additional Set**
Noisy (no preprocessing)	7.0 %	12.1 %
DNN (Kaldi)	10.8 %	15.4 %
DNN (Keras)	9.6 %	13.7 %

2.4. Oracle experiment using the STFT-phase of the clean signal for synthesis

So far, we used the STFT-phase of the noisy signal to transform back the enhanced spectra to the time-domain. To analyze the influence of the phase-information in such a DNN-based speech enhancement method we run an oracle experiment where we use the phase of the clean signal instead. Referring to figure 1 now we apply the lower right block $\arg(\cdot)$ on the STFT of the corresponding clean signal $x_{\text{clean}}[n]$ instead $x_{\text{noisy}}[n]$. The PESQ mean values are listed in table 5 to compare the speech quality with noisy and with oracle-clean phase.

Table 5: *Comparison of the influence of the STFT-phase on PESQ values in DNN-based speech enhancement*

DNN	**Aurora-4**	**Additional Set**
Keras /w noisy-phase	2.615	2.429
Keras /w clean-phase	2.948	2.736

For both sets we observe an improvement of about 0.3 in PESQ on average over all utterances due to applying the STFT-phase of the clean signal. This implies that the modification of the noisy phase has the potential of further improving the speech quality with DNN based speech enhancement methods that already provide a good estimation of the clean magnitude spectrum.

Of course in real applications the phase information of the clean signal is not available and an estimation of the clean phase has to be made. It is a well known fact that such phase estimations are far more challenging than magnitude estimation due to the unpleasant mathematic nature of the phase. As stated in the introduction, there already exist a few methods to modify the phase in the STFT-domain. As a new aspect we focus on an alternative approach that is based on a pitch synchronous analysis and synthesis in voiced speech. We analyze the influence of the magnitude spectrum and the phase spectrum in such a domain.

3. Oracle experiments based on a pitch synchronous analysis and synthesis

Most DNN-based noise reduction methods work in the STFT-domain and are applied on the entire speech signal with a constant window size regardless whether a segment contains voiced or voiceless speech. However, it is well known that voiced segments greatly contribute to speech quality. Therefore, we designed a set of oracle experiments to evaluate and quantify the effects of modifying the magnitude and phase spectrum in voiced segments based on a pitch synchronous analysis and synthesis. The modification is applied as a post processing on the enhanced signal after the DNN based modification of the magnitude spectrum. Especially, we want to analyze whether it is advisable to focus on a magnitude-only or on a phase-only modification or on the joint improvement of both.

3.1. Description of the experiment

The processing scheme of the oracle experiments is shown in figure 2 containing several switches to visualize different processing options. First, a noisy signal is enhanced with the DNN based processing shown in figure 1. We use the Keras DNN that provides better quality. We perform a detection of voiced segments either on the enhanced signal or as an oracle experiment on the clean signal. The detection method is a modified version of [20]. It is based on a first rough estimation of the pitch with a cepstral analysis technique and a more precise determination of the pitch periods based on a correlation analysis of several pitch periods. As output we get the sample positions of each pitch period and a correlation value between 0 and 1 defining the similarity of succeeding pitch periods. The voiced/unvoiced detection is done with a predefined correlation threshold. The length of the detected voiced segment increases for a lower threshold. We look at the two correlation thresholds of 0.55 and 0.7 to compare the influence of detecting more and longer segments versus the opposite case.

A DFT is applied on the detected pitch periods with a transformation length equal to the period length. We analyze each single period of an almost periodic signal. Therefore the error due to the leakage effect of the DFT is very low. The output values of the DFT approximately reflect the short term magnitude and phase spectrum at the fundamental frequency f_0 and its harmonics. Alternatively, the same periods of the clean signal are analyzed as an oracle experiment. Thus, we can look at all possible combinations of the magnitude and the phase spectra taken from the enhanced or from the clean signal. These spectra are transformed back to the time domain. During voiced segments the enhanced signal is substituted by the resynthesized signal. Using magnitude and phase of the clean signal results in the complete substitution of the enhanced signal by the clean signal.

3.2. Interpretation of results

It is noteworthy that the sole substitution of the noisy phase spectra by their clean counterpart is quite insignificant with a mean PESQ improvement less than 0.05, as listed in table 6. At first sight, this seems to be contradictory to the results shown in table 5. However, the type of analysis and synthesis is quite different at both points. In particular, the inconsistency of the STFT plays a major roll using the overlap add synthesis scheme. The effect of adding overlapping resynthesized speech segments largely contributes to the observed improvement shown in table 5.

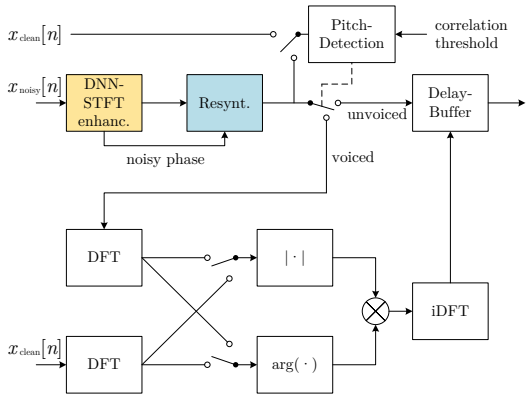

Figure 2: *Processing scheme of the oracle experiments*

Table 6: *PESQ improvement using pitch synchronous oracle-replacement of magnitude and phase spectra. 'C.T.': correlation threshold; 'M': magnitude; 'P': phase*

C.T.	pitch-detec.	param.	Aurora-4	Add. Set
0.70	DNN enh.	M+P	0.079	0.088
0.70	DNN enh.	M	0.049	0.060
0.70	DNN enh.	P	0.016	0.015
0.70	clean sig.	M+P	0.205	0.214
0.70	clean sig.	M	0.105	0.123
0.70	clean sig.	P	0.033	0.032
0.55	DNN enh.	M+P	0.122	0.132
0.55	DNN enh.	M	0.063	0.079
0.55	DNN enh.	P	0.024	0.022
0.55	clean sig.	M+P	0.313	0.322
0.55	clean sig.	M	0.140	0.166
0.55	clean sig.	P	0.047	0.046

Substituting only the magnitude DFT spectra of the noisy signal leads to a higher PESQ improvement than the substitution of the phase only. Our experiments with the oracle pitch detection using the clean signal and with the different correlation thresholds state that the potential of such methods depends on the quality of the pitch detection. If applying the pitch detection on the clean signal instead on the DNN enhanced signal, the mean PESQ values approximately double and reach values between 0.105 and 0.166. Obviously, the pitch detection has a problem to detect voiced segments in the enhanced signal. These parts of the voiced speech that are not detected seem to carry the potential for a further quality improvement. Therefore, a further improvement of the pitch detection algorithm is required for a pitch synchronous speech enhancement. Finally, the experiments state that the joint modification of magnitude and phase spectra achieves far better results than a magnitude-only modification or a phase-only modification. In the best case of our oracle experiments this can lead to an improvement of 0.3 on top of the already achieved DNN-STFT-enhancement.

4. Conclusion and outlook

In this work we set up a DNN-based speech enhancement method that is based on a mapping of the noisy magnitude spectra to the corresponding clean spectra. We evaluated this method in terms of improving the PESQ value as measure for the speech quality and in terms of improving the recognition rate of an ASR system. Furthermore, we showed in this context that the STFT-phase holds potential for further improvement. Based on a pitch synchronous post processing of voiced segments we run a set of oracle experiments. We identified the simultaneous modification of magnitude and phase spectra as the best choice to further improve the speech quality. Modifying only the magnitude or only the phase clearly has less potential. By applying a pitch synchronous DFT analysis we reduce the complexity in comparison to a STFT analysis with a fixed window size. Furthermore, we hope that these pitch synchronous spectra make it easier to estimate not only an enhanced version of the magnitude spectrum but also an enhanced version of the phase spectrum by applying a deep neural network. Our future work will focus on this topic.

5. References

[1] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 1, pp. 7–19, 2015.

[2] Y. Xu, J. Du, Z. Huang, L.-R. Dai, and H. Lee, "Multi-objective learning and mask-based post-processing for deep neural network based speech enhancement," *INTERSPEECH 2015*, pp. 1508–1512, 2015.

[3] P. Mowlaee, R. Saeidi, and Y. Stylianou, "Advances in phase-aware signal processing in speech communication," *Speech Communication*, vol. 81, pp. 1–29, 2016.

[4] T. Gerkmann, M. Krawczyk, and R. Rehr, "Phase estimation in speech enhancement - unimportant, important, or impossible?" *2012 IEEE 27th Convention of Electrical & Electronics Engineers in Israel (IEEEI 2012)*, pp. 1–5, 2012.

[5] L. D. Alsteris and K. K. Paliwal, "ASR on speech reconstructed from short-time fourier phase spectra," *Proc. International Conf. Spoken Language Processing*, 2004.

[6] M. Krawczyk and T. Gerkmann, "STFT phase reconstruction in voiced speech for an improved single-channel speech enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 1931–1940, 2014.

[7] K. Li, B. Wu, and C.-H. Lee, "An iterative phase recovery framework with phase mask for spectral mapping with an application to speech enhancement," *Interspeech 2016*, pp. 3773–3777, 2016.

[8] D. W. Griffin and J. S. Lim, "Signal estimation from modified short-time fourier transform," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 2, pp. 236–243, 1984.

[9] ITU-T, "Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," *ITU-T Recommendation P.862 (02/2001)*, 2001.

[10] ——, "Mapping function for transforming P.862 raw result scores to MOS-LQO," *ITU-T Recommendation P.862.1 (11/2003)*, 2003.

[11] N. Parihar, J. Picone, D. Pearce, and H. G. Hirsch, "Performance analysis of the Aurora large vocabulary baseline system," *2004 12th European Signal Processing Conference*, vol. 12, pp. 553–556, 2004.

[12] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The kaldi speech recognition toolkit," *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*, 2011.

[13] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "Dynamic noise aware training for speech enhancement based on deep neural networks," *INTERSPEECH 2014*, pp. 2670–2674, 2014.

[14] H.-G. Hirsch and H. Finster, "A new approach for the adaptation of HMMs to reverberation and background noise," *Speech Communication*, vol. 50, no. 3, pp. 244–263, 2008.

[15] N. Parihar and J. Picone, "Aurora working group: DSR front end LVCSR evaluation AU/384/02," 2002.

[16] D. B. Paul and J. M. Baker, "The design for the Wall Street Journal-based CSR corpus," *Proceedings of the Workshop on Speech and Natural Language*, pp. 357–362, 1992.

[17] X. Aubert, C. Dugast, H. Ney, and V. Steinbiss, "Large vocabulary continuous speech recognition of Wall Street Journal data," *Acoustics, Speech, and Signal Processing, 1994. ICASSP-94., 1994 IEEE International Conference on*, vol. ii, pp. II/129–II/132 vol.2, 1992.

[18] H.-G. Hirsch, "Extraction of robust features by combining noise reduction and FDLP for the recognition of noisy speech signals in hands-free mode," *The REVERB Workshop (held in conjunction with ICASSP 2014 and HSCMA 2014)*, 2014.

[19] C. Breithaupt, T. Gerkmann, and R. Martin, "Cepstral smoothing of spectral filter gains for speech enhancement without musical noise," *IEEE Signal Processing Letters*, vol. 14, no. 12, pp. 1036–1039, 2007.

[20] H.-G. Hirsch and F. Kremer, "Recognition of noisy speech by starting the likelihood calculation at voiced segments," *Speech Communication: 11. ITG Symposium*, pp. 1–4, 2014.