Single-channel noise suppression based on a statistical source-model for speech

Niklas Harlander, Thomas Rohdenburg and Volker Hohmann
Medizinische Physik, Carl-von-Ossietzky Universität Oldenburg, Germany

We propose a single-channel noise suppression scheme based on a statistical source-model for speech. The scheme is adapted from Ephraim and Malah (1984) and Tchorz and Kollmeier (2003) and aims at improving short-time signal-to-noise ratio (SNR) estimates in different frequency subbands by learning and classifying auditory-model based speech signal features. First, the speech signal is transformed into so-called Amplitude-Modulation-Spectrograms (AMS) firstly described in Kollmeier and Koch (1994), which include information of both center frequencies and modulation frequencies within 32-ms analysis frames. Second, the short-time subband SNR is estimated from the AMS patterns by a neural network, which was trained based on a large speech database. A second neural net obtains final SNR estimates from (i) the AMS-based SNR estimates by Tchorz and Kollmeier (2003), and (ii) the estimates derived from the traditional approach by Ephraim and Malah (1984). The final SNR estimates can be used to steer a Wiener filter for noise suppression. Experimental results indicate a reasonable SNR-estimation accuracy.

INTRODUCTION

Noise reduction remains an important issue largely due to the wide field of applications of speech signal processing, e.g., in hearing aids, mobile phones and speech recognition systems, to name only some. A crucial component of noise reduction schemes is the estimation of the noise power spectrum. Whereas efficient schemes exist for the special case of stationary noise conditions the general case of fluctuating noise conditions as encountered in many realistic sound fields is still a challenging problem. The aim of the study is to enhance single-channel noise suppression based on improved SNR estimation. This is realized by the combination of an Ephraim/Malah-type scheme and the neurophysiologically motivated scheme by Tchorz and Kollmeier (2003) Based on a brief description of both schemes, the novel SNR estimation algorithm will be introduced and evaluated.

Algorithm by Ephraim and Malah

The algorithm of Ephraim and Malah (1984) is a well-known *minimum mean square error (MMSE) short time spectral attenuation (STSA)* technique that aims at suppression of stationary noise. This approach is based on modelling speech and noise spectral components as statistically independent Gaussian random variables. In the following the most important characteristics shall be summarised. For the algorithm to work properly an estimation of the noise floor is required. Several estimation methods exist for this purpose. In this study the following two methods were tested:

Auditory signal processing in hearing-impaired listeners. 1st International Symposium on Auditory and Audiological Research (ISAAR 2007). T. Dau, J. M. Buchholz, J. M. Harte, T. U. Christiansen (Eds.).
ISBN: 87-990013-1-4. Print: Centertryk A/S.

- Voice Activity Detector (VAD) described in Marzinzik and Kollmeier (2002) is based on an adaptive detection of the power envelope minima of the broadband-, lowpass- and highpass-filtered signal.
- Detection of spectral minima such as in R. Martin (1994).

The efficient reduction of stationary noise is basically due to the combination of the *a posteriori* SNR, which is an instantaneous SNR estimate and the a *priori* SNR. The a *priori* SNR is the averaged SNR based on previous estimates. In this way a noticeable suppression of "musical tones" is attained. For this type of algorithm an enhancement of speech quality but no enhancement of speech intelligibility was found empirically by Marzinzik and Kollmeier (2001).

Algorithm by Tchorz and Kollmeier

The algorithm by Tchorz and Kollmeier (2003) is particularly interesting because of its neurophysiological and psychoacoustical motivation. It uses the so-called Amplitude-Modulation-Spectrogram (AMS) by Kollmeier and Koch (1994), which conveys information of both center frequencies and modulation frequencies. Those AMS patterns are used as features and are classified by a neural network (NN). This kind of pattern recognition leads to across-channel processing and results in an analysis and classification of spectro-temporal information of speech and typical noise. In addition there is neither need of a VAD nor the need of assumptions about stationarity of the noise. The algorithm can be characterized by a training and the subsequent testing phase as depicted in Fig. 1.

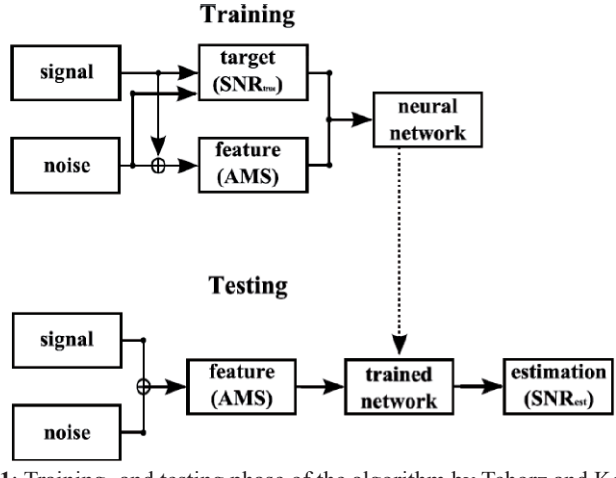

Fig. 1: Training- and testing phase of the algorithm by Tchorz and Kollmeier (2003).

In the training-phase (upper row of processing blocks in Fig. 1) the signal and the noise are separately available. The input parameters for the neural network (NN) are a target vector and a feature matrix. The target vector consists of the “true” SNR per frequency band which was calculated by the power density spectrum of the signal divided by the power density spectrum of the noise in third octave bands. The feature extraction and the parameters of the NN are described below. In the training phase the NN

is trained on a large training corpus to optimally estimate the target SNR from the feature matrix. In the testing phase (lower row of processing blocks in Fig. 1), the trained network estimates the SNR from the feature matrix. Those estimates can then be subsequently used to steer a Wiener filter and to suppress noise.

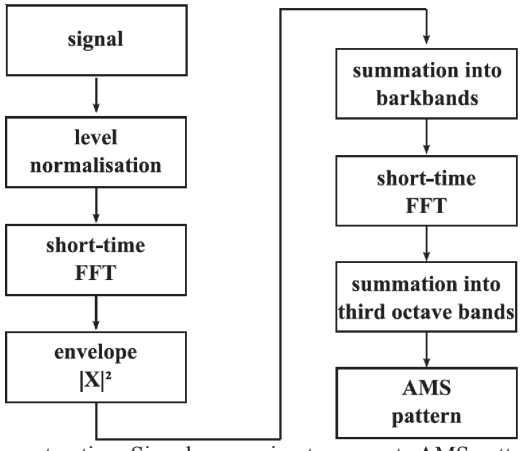

Fig. 2: Feature extraction: Signal processing to generate AMS patterns.

Feature extraction is shown in Fig. 2. First, the signal is sampled at 16 kHz. It is subdivided into segments of 64 samples with an overlap of 60 samples. Each segment is divided by the absolute maximum of the segment so that the level of the signal is normalised, and then subjected to a short-time Fast Fourier Transformation (FFT). The envelope in each of 15 frequency bands is derived by non-overlapping summation of squared FFT-magnitudes across critical bands (ERB resolution) and then subjected to another short-time FFT. For this, segments of 128 samples with an overlap of 64 samples are used. Modulation intensities are summed up across non-overlapping third-octave bands, resulting in a total of 15 modulation bands. In summary, each 32-ms segment of the input signal is represented by one AMS pattern that contains modulation intensities in 15 modulation bands in each of 15 spectral bands. Each AMS pattern forms the feature matrix used by the NN to estimate the SNR of the corresponding signal segment.

Examples of AMS patterns are plotted in Fig. 3. Voiced speech (first panel) and pink noise (second panel) clearly show different characteristics. The AMS pattern of the voiced speech shows a certain structure presumably due to the periodicity of the fundamental frequency and formants, respectively. In contrast the AMS pattern of the noise exhibits no specific sub-structure. Only a spectral tilt is noticeable at higher frequencies because the energy of pink noise falls off by 3 dB per octave.

For classification and estimation, respectively, a feed-forward, multi layer perceptron by I. Nabney (2004) is used. Its parameters are depicted in Table 1. 225 input neurons were used which corresponds to the resolution of the AMS patterns (15x15). The 15 output neurons are used to estimate the SNR per frequency channel.

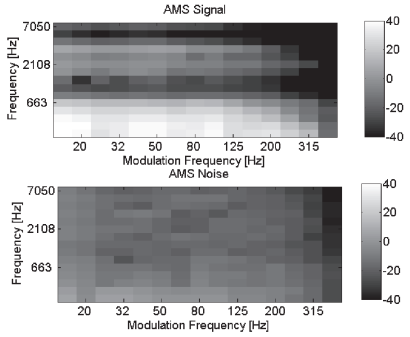

Fig. 3: AMS pattern of a 32 ms segment of voiced speech (first panel) and pink noise (second panel). Bright and dark areas indicate high and low intensities, respectively.

Parameters	***Size***
input neurons (AMS pattern 15x15)	225
hidden neurons	160
output neurons (SNR in 15 bands)	15
activation function of hidden neurons	*tanh*
activation function of output neurons	*softmax* [0,1]
repetitions in the training phase	100

Table 1: Set of parameters for the neural network.

ALGORITHM

The algorithm used in this study combines the two independent algorithms of Ephraim and Malah Ephraim and Malah (1984) and Tchorz and Kollmeier (2003). Both algorithms separately estimate the SNR in 15 subbands as described above. Then a second neural network is trained that provides a final SNR estimate (SNR_{EMTCH}) from the Ephraim and Malah estimates (SNR_{EMpost} , SNR_{EMapri}) and the estimate based on the algorithm of Tchorz (SNR_{TCH}). The block diagram of the algorithm is depicted in Fig. 4.

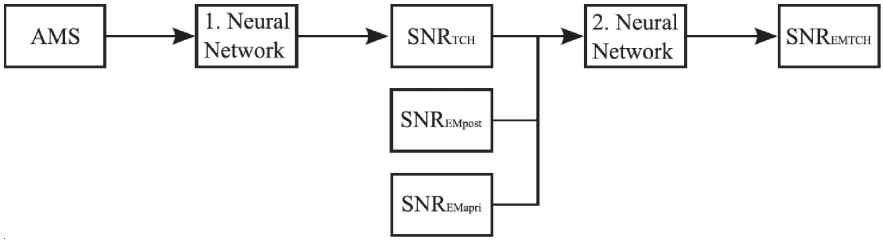

Fig. 4: Algorithm to estimate the SNR from the superposition of speech and noise. The scheme combines the estimates delivered by the algorithm Ephraim and Malah (1984) and Tchorz and Kollmeier (2003).

RESULTS

The training and the test data which were presented to the neural networks were taken from the "PhonDat" (1995) corpus. The noisy speech was artificially mixed with 16 noise types taken from various databases with a global SNR of 5 dB, which corresponds to limited segmental range of SNR's from -10 up to 20 dB. In total, 16 minutes speech material was used for training and 10 minutes for testing. To obtain a quantitative measure of the estimation accuracy the mean deviation D is calculated by

$$D = \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{15} | e_{ij} - a_{ij} | \qquad \text{(Eq. 1)}$$

where a_{ij}, is the "true" SNR in segment i and frequency band j, e_{ij}, is the corresponding estimated SNR and N is the number of signal segments processed. The mean deviations of the combination of the algorithms are compared to the results from the single algorithms. The aim of these experiments was to examine if the combination of the algorithms enhances the SNR estimation. It was not intended to find the optimal parameter set for each algorithm but to verify the improved SNR estimation. Fig. 5 shows the results using the VAD for the Ephraim/Malah-type algorithm (*EM*). The results for the algorithm after Tchorz are denoted by *TCH*.

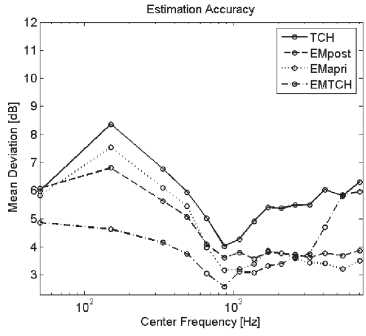

Fig. 5: Mean Deviation. Noise floor estimation: VAD by Marzinzik and Kollmeier (2002).

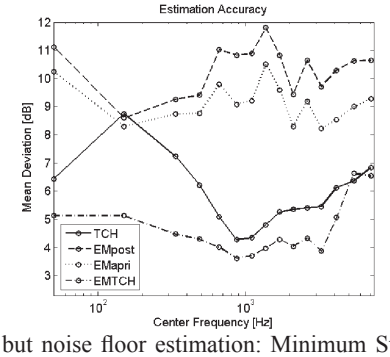

Fig. 6: Same as Fig. 5 but noise floor estimation: Minimum Statistics by R. Martin (1994).

The combination of the algorithms shows the best performance for frequencies below 4 kHz. The estimation is getting worse towards higher frequencies, probably because the NN was insufficiently trained to disregard the worse estimates from *TCH* at these frequencies. Fig. 6 shows the corresponding results for the Minimum Statistics approach by Martin (1994). Surely the parameter set of this algorithm could be optimized but still there is clearly an enhancement by combining both estimates.

The next experiment shall underline this thesis. A single speech file was tested and the narrow-band estimations in one frequency channel out of 15 are depicted in Fig. 7. It is apparent that the combination leads to an enhanced SNR estimation. If one algorithm overestimates the SNR while the other underestimates it, the combination finds a balance between both estimates.

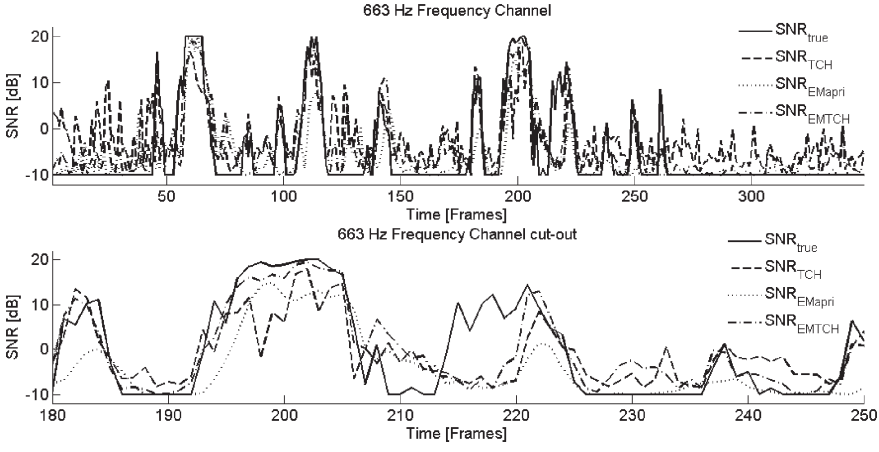

Fig. 7: Narrow-band estimation. One frequency channels out of 15 (first panel) plus cut-out (second panel).

Fig. 8 shows the estimated SNR vs. the true SNR accumulated across all frequency bands. The number of occurrences of a combination of true and estimated SNR is plotted on a greyscale. Apart from the obvious spread that was quantified above in Figs. 5 and 6, a trend to underestimating the SNR is observed.

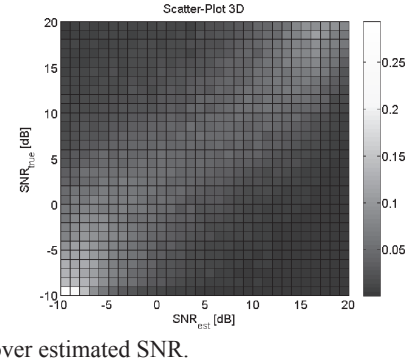

Fig. 8: True SNR over estimated SNR.

Finally the hit rate was calculated by subtracting the absolute value of the true SNR from the estimated SNR and counting the number of occurrences of values smaller than one, three and five dB. Fig. 9 shows the hit rate in percent. 72 % of the data is estimated with an error below 5 dB.

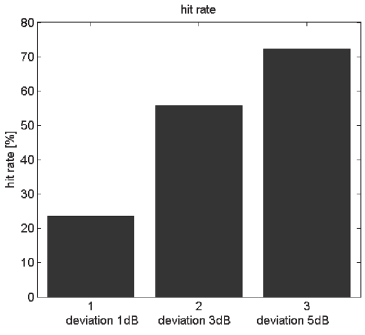

Fig. 9: Hit rate divided into sections of one, three and five *dB* deviation in the SNR.

CONCLUSION

A single-channel noise suppression scheme was proposed based on a statistical source-model for speech that combines the approaches of Ephraim and Malah (1984) and Tchorz and Kollmeier (2003). The combination was found to improve the SNR estimates compared to the estimates from each of the two algorithms alone. Although not tested in this study, the improved SNR estimates may potentially lead to an enhancement of single-channel noise suppression schemes. The aim of this study was not finding the optimal parameter set for the state-of-the-art algorithms but rather to exhibit a reasonable improvement of the SNR estimation by the combination of the algorithms. Nevertheless, it would be interesting to review the results using an optimized set of parameters of the contributing algorithms Ephraim and Malah (1984) and Tchorz and Kollmeier (2003).

REFERENCES

Ephraim, Y., and Malah, D. (**1984**). "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator". IEEE Signal Proc. Letters, ASSP-**32**,1109 – 1121.

Kollmeier, B., and Koch, R. (**1994**). „Speech enhancement based on physiological and psychoacoustical models of modulation perception and binaural interaction." J. Acoust. Soc. Am., **95**, 1593 – 1602.

Martin, R. (**1994**). "Spectral subtraction based on minimum statistics". Proc. European Signal Processing Conference, **1**, 1182 – 1185.

Marzinzik, M., and Kollmeier, B. (**2002**). "Speech pause detection for noise spectrum estimation by tracking power envelope dynamics." IEEE Transactions on Speech and Audio Processing ASSP-**10**, 109– 118.

Marzinzik, M., and Kollmeier, B. (**2001**). "A review of the Ephraim-Malah noise reduc-

tion algorithms," Zeitschrift für Audiologie/Audiological Acoustics ASSP-**40**, 4–15.

Nabney, I. T. (**2004**). "NETLAB. Algorithms for Pattern Recognition." Springer Verlag, Berlin, 3rd edition.

Universität München, Institut für Phonetik und sprachliche Kommunikation (ipsk). (**1995**). BAS (Bayerisches Archiv für Sprachsignale) Phondat 1 - pd1 und Phondat 2 - pd2 (Phonitatrische Datenbank). http://www.bas.uni-muenchen.de/Bas.

Tchorz, J., and Kollmeier, B. (**2003**). "SNR Estimation Based on Amplitude Modulation Analysis With Applications to Noise Suppression," IEEE Trans. Speech and Audio Processing, **11**, 184-192.