

# Automatic classification of the acoustical situation using amplitude modulation spectrograms

Jürgen Tchorz and Birger Kollmeier

AG Medizinische Physik, Universität Oldenburg, 26111 Oldenburg, Germany  
tch@medi.physik.uni-oldenburg.de

**Summary:** This paper describes a sound classification algorithm which automatically distinguishes between the two sound source classes „speech“ and „noise“. In situations with both speech and noise at the same time, the algorithm can predict the current SNR to some extent. The approach bases on so-called Amplitude Modulation Spectrograms (AMS). In AMS patterns, both spectral and temporal properties of short signal segments (32ms) are represented, which is motivated by neurophysiological findings on auditory processing in mammals. Artificial neural networks are trained on a large number of AMS patterns from speech and noise and are then used to classify new patterns.

## INTRODUCTION

Digital processing in modern hearing instruments allows the implementation of a wide range of sound processing schemes. Monaural and binaural noise suppression can enhance speech quality and speech intelligibility in adverse acoustical environments. However, the performance of most of these algorithms is strongly dependent on a proper and reliable classification of the acoustical situation in terms of a speech/noise decision. If both speech and noise are present, an estimate of the signal-to-noise ratio is desired. Furthermore, the classification of the monitoring algorithm should work on short time scales, because most noise suppression algorithms need an exact detection of speech pauses for good performance. The classification algorithm which is presented in this paper bases on so-called amplitude modulation spectrograms (AMS). Its basic idea is that both spectral and temporal information of the signal is used to attain a separation between „acoustical objects“ within the signal. The AMS approach is motivated by neurophysiological experiments on periodicity coding in the auditory cortex of mammals, where neurons tuned to different center frequencies were found to be organized almost orthogonal to neurons which are tuned to different modulation frequencies (1). Kollmeier and Koch (2) implemented a binaural noise suppression

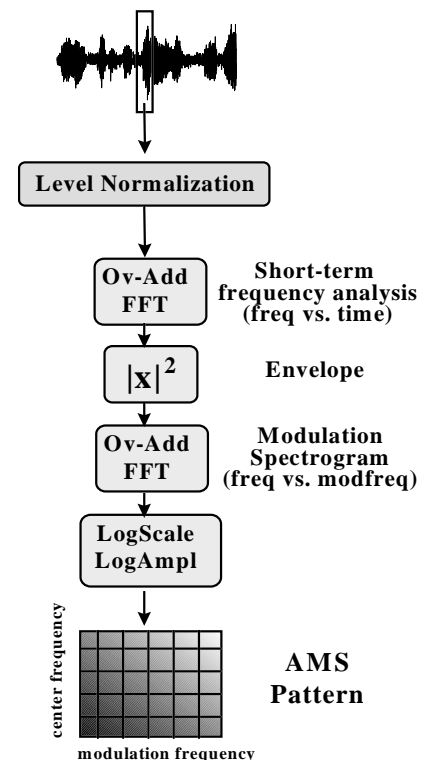


Figure 1: Signal processing steps for AMS pattern generation.

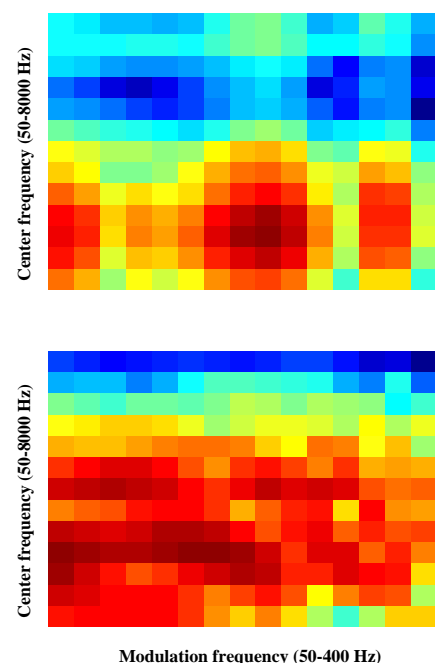
scheme which bases on AMS sound representation. They could demonstrate a benefit in terms of speech intelligibility, in comparison to unprocessed speech.

## GENERATING AMS PATTERNS

Figure 1 shows the processing steps which are performed to generate AMS patterns. First, the input signal is long-term level adjusted, i.e., changes in the overall level are compensated for, whereas short-term level differences (e.g., those between successive phonemes) are maintained to serve as additional cues for classification. This level adjustment is realized by dividing the input signal by its 2 Hz-low pass filtered running RMS value. The level-adjusted signal is then subdivided into overlapping segments of 4.0 ms duration with a progression of 0.25 ms for each new segment. Each segment is multiplied with a Hanning window and padded with zeros to obtain a frame of 128 samples which is transformed with a FFT into a complex spectrum. The resulting 64 complex samples are considered as a function of time, i.e., as bandpass-filtered complex time signal. Their respective envelopes are extracted by squaring. This envelope signal is again segmented into overlapping segments of 128 samples (32ms) with an overlap of 64 samples. A further FFT is computed and supplies a modulation spectrum in each frequency channel. By an appropriate summation of neighboring FFT bins, both axes are scaled logarithmically with a resolution of 15 channels for center frequency (100-7300 Hz) and 15 channels for modulation frequency (50-400 Hz). In a last processing step, the amplitude range is log-compressed. Examples for AMS patterns can be seen in Fig. 2. The AMS pattern on top was generated from a voiced speech portion. The periodicity at the fundamental frequency (approx. 250 Hz) is represented in each center frequency band. The AMS pattern on the bottom was generated from speech simulating noise. The typical spectral tilt can be seen, but no structure across modulation frequencies.

## NEURAL NETWORK TRAINING AND CLASSIFICATION EXPERIMENTS

Speech/noise classification of AMS patterns is performed with a standard feed-forward neural network. It consists of an input layer with 225 (15x15) neurons, a hidden layer with 20 neurons, and an output layer with 1 output neuron. The target activity of the output neuron for training is set to 0.95 or 0.05 for AMS patterns generated from speech or noise, respectively. For training, 495 sec of speech from 28 different talkers and 495 sec of noise from 14 different noise sources were transformed into AMS patterns (61896 AMS training patterns in total). After training,



**Figure 2:** AMS patterns generated from speech (top) and from noise (bottom). Each AMS pattern represents a 32 ms-analysis frame of the input signal.

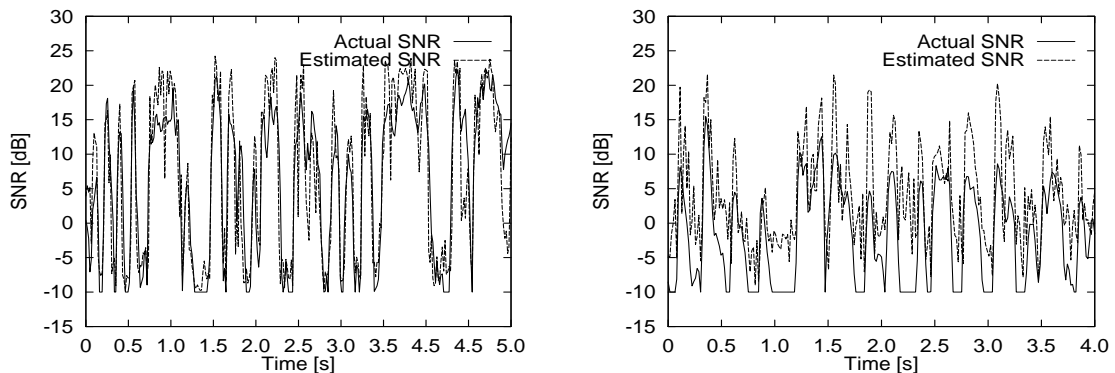
AMS patterns from "unknown" signal segments are classified with an output neuron activity threshold of 0.5: If the network responds to an AMS pattern with an output neuron activity above 0.5, the segment is classified as "speech", otherwise as "noise". Classification results on different databases are shown in Table 1. For speech, the best classification results were gained when the test data

**TABLE 1:** The influence of different databases for training and testing on correct speech/noise classification

	Speech		Noise	
	Database	Recognition Rate [%]	Database	Recognition Rate [%]
Training	phondat	98.4	Various, partly alife	99.5
Testing	phondat	95.7	but not noisex	
	timit	91.7	alife	92.7
	zifkom	93.6	noisex	96.8

originated from the same database as the training material (PHONDAT). Degraded performance for TIMIT-data can be explained by different long-term spectra of these databases due to differences in the recording conditions.

In situations with both speech and noise present, an estimation of the present signal-to-noise ratio (SNR) is desired. To achieve this, the network is trained with AMS patterns generated from mixtures of speech and noise. The target activity then depends on the "local" SNR of the according AMS analysis frame. The SNR range from 25 to -10 dB is linearly transformed to target activities between 0.95 or 0.05. SNRs above 25 dB and below -10 dB are transformed to 0.95 and 0.05, respectively. The training AMS patterns were generated from a mixture of the training material that was used for the pure speech/pure noise classification described above. After training, the output neuron activity of the network when presenting "unknown" AMS patterns serves as estimate for the present SNR. In Fig. 3, two examples of SNR estimations are illustrated. The solid lines show the actual SNR which was determined before adding speech and noise, the dotted lines show the estimate of the SNR (after re-transforming output neuron activity to SNR). On the left, the input signal was a mixture of speech and car noise. In that condition, the algorithm can predict the present SNR with satisfactory accuracy. On the right, the input signal was a mixture of speech and cafeteria noise. Here, the estimated SNR is higher than the actual SNR most of the time. This might be due to the "speech-like" characteristic of cafeteria noise. Automatic SNR estimation can be described quantitatively by measuring the mean deviation between the actual and the



**Figure 3:** SNR estimation for two different input signals. Left: mixture of speech and car noise. Right: mixture of speech and cafeteria babble. The solid lines show the actual local SNR which was measured before adding the signals. The dotted lines show the estimation of the local SNR provided by the algorithm.

**Table 2:** Database configurations for training and testing. See Fig. 4

Confi- guration	Training material		Test material	
	speech	noise	speech	noise
1	phondat	various	phondat	noisex
2	phondat	alife	phondat	noisex
3	phondat	various	phondat	alife
4	phondat	alife	phondat	various
5	phondat	alife	phondat	Cosmea-M
6	phondat	alife	Cosmea-M	Cosmea-M

4. The different test configurations are explained in Table 2. For most configurations, the mean deviation between measured and estimated SNR is about 4.5 dB for a frame-by-frame estimation (each AMS pattern provides an independent estimate), and in the range of 2.5 dB for low pass filtering with  $f_{\text{cut-off}}=1$  Hz, i.e., the algorithm adapts to a new SNR within about 500 ms. One exception is configuration 6. Here, test samples for both speech and noise were recorded with the same microphone (Siemens Cosmea-M) which were placed into the ears of the same person (using ear moulds). Cosmea-M samples were not included in the training data. These two data bases spectrally differ significantly from each other. Thus, better results can be expected for taking material from the same database for training and testing, for both speech and noise.

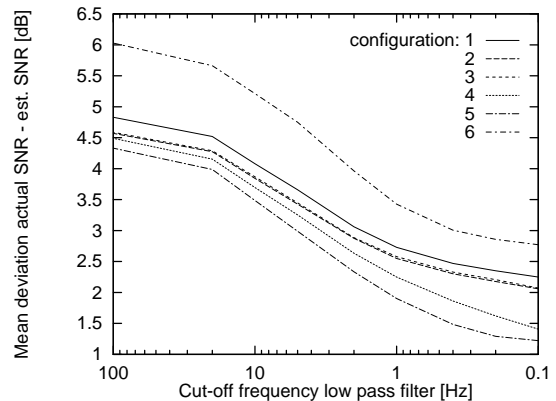
## DISCUSSION AND CONCLUSION

It was demonstrated that the combination of spectral and temporal information of acoustic signals in AMS patterns can be exploited for automatic classification of the acoustical situation. The precision of SNR estimation in situations where both speech and noise are present is dependent on the background noise. The performance might be improved by extending the amount of training data and modifying the SNR - target activity transformation function. A sub band SNR estimation (by increasing the number of output neurons) would allow for noise reduction by attenuation of bands with bad SNR and will be implemented and tested in further experiments.

## REFERENCES

- (1) Langner, G. *Hear. Res.* **60**, 115-142 (1992)
- (2) Kollmeier, B. and Koch, R. *J. Acoust. Soc. Am.* **95**, 1593-1602 (1994)

estimated SNR for each AMS analysis frame. Depending on the application, outliers and estimation errors can be reduced by temporal smoothing (filtering) of successive SNR estimates. The mean deviation as a function of low pass filter frequency for a range of different database configurations is plotted in Fig.



**Figure 4:** Mean deviation between the actual SNR and the estimated SNR as a function of lowpass filtering of successive SNR values. The different configurations are explained in Table 2.