USING AMPLITUDE MODULATION INFORMATION FOR SOUND CLASSIFICATION

JÜRGEN TCHORZ AND BIRGER KOLLMEIER
AG Medizinische Physik, Universität Oldenburg, 26111 Oldenburg, Germany
E-mail: tch@medi.physik.uni-oldenburg.de

1 Introduction

Digital processing in modern hearing aids 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 these algorithms is strongly dependent on a proper and reliable classification of the acoustical situation. Noise suppression techniques like spectral subtraction, for example, are strongly dependent on a fast and proper estimate of the present noise level. Thus, an important task of the classification algorithm is to decide whether speech or noise is present. 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 scheme which bases on AMS sound representation. They could demonstrate a benefit in terms of speech intelligibility, in comparison to unprocessed speech.

2 Generating AMS patterns

In a first processing step, 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

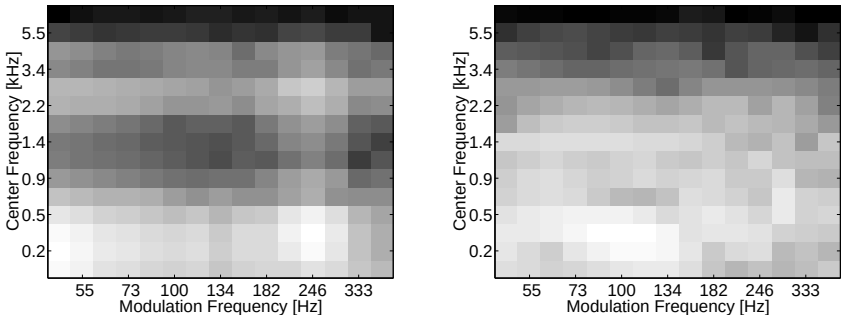

Figure 1. AMS pattern generated from voiced speech (left) and speech simulating noise (right). Bright and dark shading indicates high and low amplitude, respectively. Each AMS pattern represents a 32-ms portion of the input signal.

additional cues for classification. This level adjustment is realized by dividing the input signal by its 2 Hz-low pass filtered rms function. 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, and transformed into a complex spectrum with a FFT. 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. 1. The left AMS pattern was generated from a voiced speech portion. The formant structure is represented, as well as the periodicity at the fundamental frequency (i.e., approx. 250 Hz) in each center frequency band. The AMS pattern on the right was generated from speech simulating noise. The typical spectral tilt can be seen, but no structure across modulation frequencies.

3 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,

Table 1. Classification results in pure speech/pure noise conditions

	database	# of AMS patterns	% correct
training data:			
speech	PHONDAT	30948	98.4
noise	various	30948	99.5
test data:			
speech	PHONDAT	10383	95.7
	ZIFKOM	9829	93.6
	TIMIT	1573	91.7
noise	NOISEX	10383	96.8
	ALIFE	2507	92.7

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 s of speech from 28 different talkers and 495 s of noise from 14 different noise sources were used, yielding 61896 AMS training patterns in total. After training, AMS 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 Tab. 1. For speech, the best classification results were gained when the test data 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.

4 SNR estimation in mixed situations

In situations where both speech and noise are present, segment classification as either "speech" or "noise" is not appropriate. Instead, 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 described in Sec. 3. After training, the output neuron activity of the network when presenting "un-

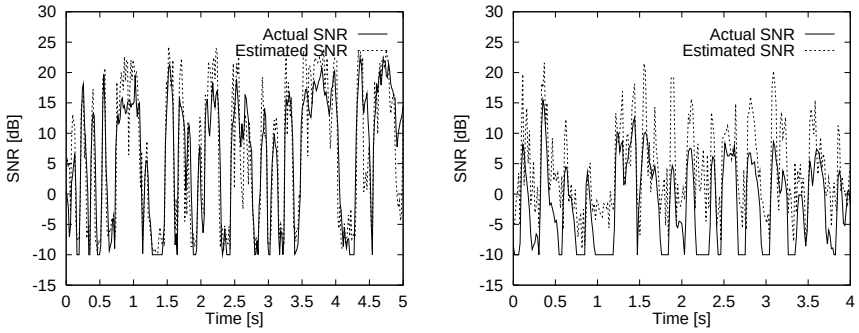

Figure 2. Estimation of SNR in car noise (left) and cafeteria noise (right).

known" AMS patterns serves as estimate for the present SNR. In Fig. 2, two examples of SNR estimations are illustrated. The solid lines show the actual SNR, 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.

5 Conclusion

It was demonstrated that the combination of spectral and temporal information of acoustic signals 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.

References

1. G. Langner, M. Sams, P. Heil, and H. Schulze, J. Comp. Physiol. A **181**, 665-676 (1997).
2. B. Kollmeier and R. Koch, J. Acoust. Soc. Am. **95**, 1593-1602 (1994).