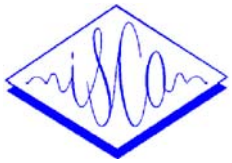

6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

SPEECH DETECTION AND SNR PREDICTION BASING ON AMPLITUDE MODULATION PATTERN RECOGNITION

Jürgen Tchorz and *Birger Kollmeier*
AG Medizinische Physik, Universität Oldenburg,
26111 Oldenburg, Germany
tch@medi.physik.uni-oldenburg.de

10.21437/Eurospeech.1999-526

ABSTRACT

A sound classification algorithm is presented which estimates the signal-to-noise ratio between speech and noise in 15 different frequency channels. The algorithm bases on the extraction of spectro-temporal features from the acoustical waveform. The approach is motivated by neurophysiological findings on periodicity coding in the auditory system of mammals. The extracted feature patterns are called Amplitude Modulation Spectrograms (AMS), as each AMS pattern contains information on both center frequencies and amplitude modulations in a short segment (32ms) of the input signal. An artificial neural network is trained on a large set of AMS patterns from mixtures of speech and noise and is then used to predict the narrow-band signal-to-noise ratio of „unknown" sounds.

Keywords: signal-to-noise ratio, classification, amplitude modulations, neural networks

1. INTRODUCTION

The automatic classification of the acoustical situation is an important topic in a range of signal processing algorithms. Noise suppression in digital hearing instruments, for example, requires a fast and reliable detection of speech pauses to effectively remove noise from disturbed signals. Classification errors typically lead to unwanted degradation of the speech signal. Humans can easily detect and classify different sound sources, and separate between speech and noise without problems. This is made possible by the interplay between the internal representation of sounds in the auditory system, and the higher processing stages in the brain which perform classification, recognition, and understanding basing on this „internal picture" of sounds. It is still unknown which are the most important features inside the acoustical waveform that allow for such impressive skills. Besides the well-known analysis of different center frequencies in the auditory system (e.g., on the basilar membrane), the representation of different *modulation* frequencies might provide important information for later processing stages. Langner and Schreiner, among others, found neurons in the inferior colliculus and auditory cortex of mammals which were tuned to certain modulation frequencies. The „peridotopical" organization of these neurons with respect to different best modulation frequencies was found to be almost orthogonal to the tonotopical organization of neurons with respect to center frequencies [1]. Thus, a two-dimensional "feature set" represents both spectral and temporal properties of the acoustical signal. Kollmeier and Koch [2] applied these findings in the field of digital signal processing and used two-dimensional, so-called Amplitude Modulation Spectrograms (AMS) in a binaural noise suppression scheme. Recently, similar kinds of feature patterns were applied to vowel segregation [3], speech enhancement [4], and broad band SNR prediction [5]. This paper presents the application of AMS patterns for narrow-band SNR prediction, which allows for the estimation of noise energy in different bands even if speech is present at the same time.

SIGNAL PROCESSING

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

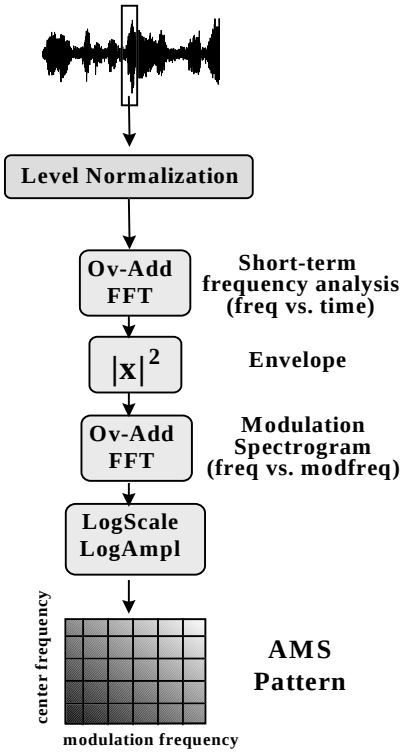

Figure 1: Signal processing steps for AMS pattern generation.

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 band pass 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. 120 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.

3. NEURAL NETWORK CLASSIFICATION

For classifying AMS patterns and estimating the narrow-band SNR of each AMS pattern, a feed-forward neural network was implemented. The net consisted of 225 input neurons (15*15, the AMS resolution of center frequencies and modulation frequencies, respectively), a hidden layer with 160 neurons, and an output layer with 15 neurons. The activity of each output neuron indicates the SNR in one of the 15 center frequency channels. For training, the narrow-band SNRs in 15 channels were measured for each AMS analysis frame of the training material prior to adding speech and noise. The measured SNR values were transformed to output neuron activities which served as target activities during training (SNRs between –10 and 30 dB were linearly transformed to activities between 0.05 and 0.95. SNRs below –10 dB and above 30 dB were assigned to activities of 0.05 and 0.95, respectively). After training, AMS patterns generated from „unknown“ sound material were presented to the network. The 15 output neuron activities that appeared for each pattern were linearly re-transformed and served as SNR estimates for the respective frequency channels.

4. SOUND MATERIAL

For training, a mixture of speech and noise with a total length of 72 min was processed and

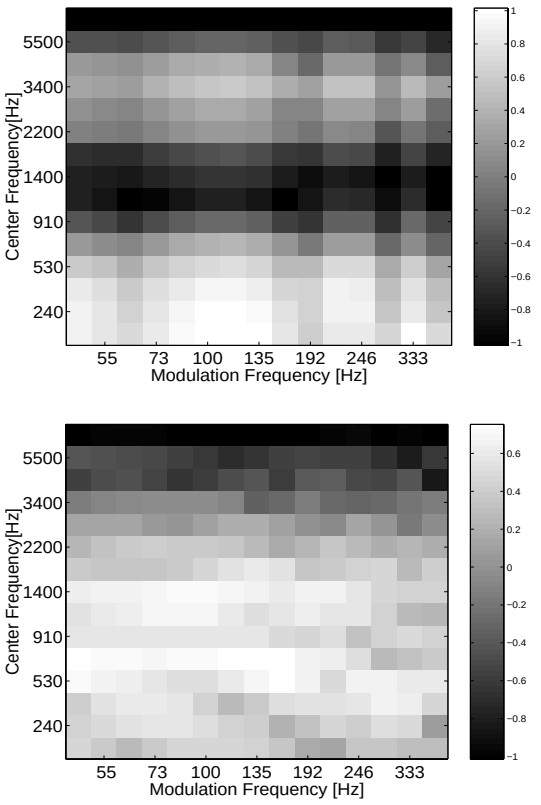

Figure 2: AMS patterns generated from speech (top) and from noise (bottom). Each AMS pattern represents a 32 ms-analysis frame of the input signal.

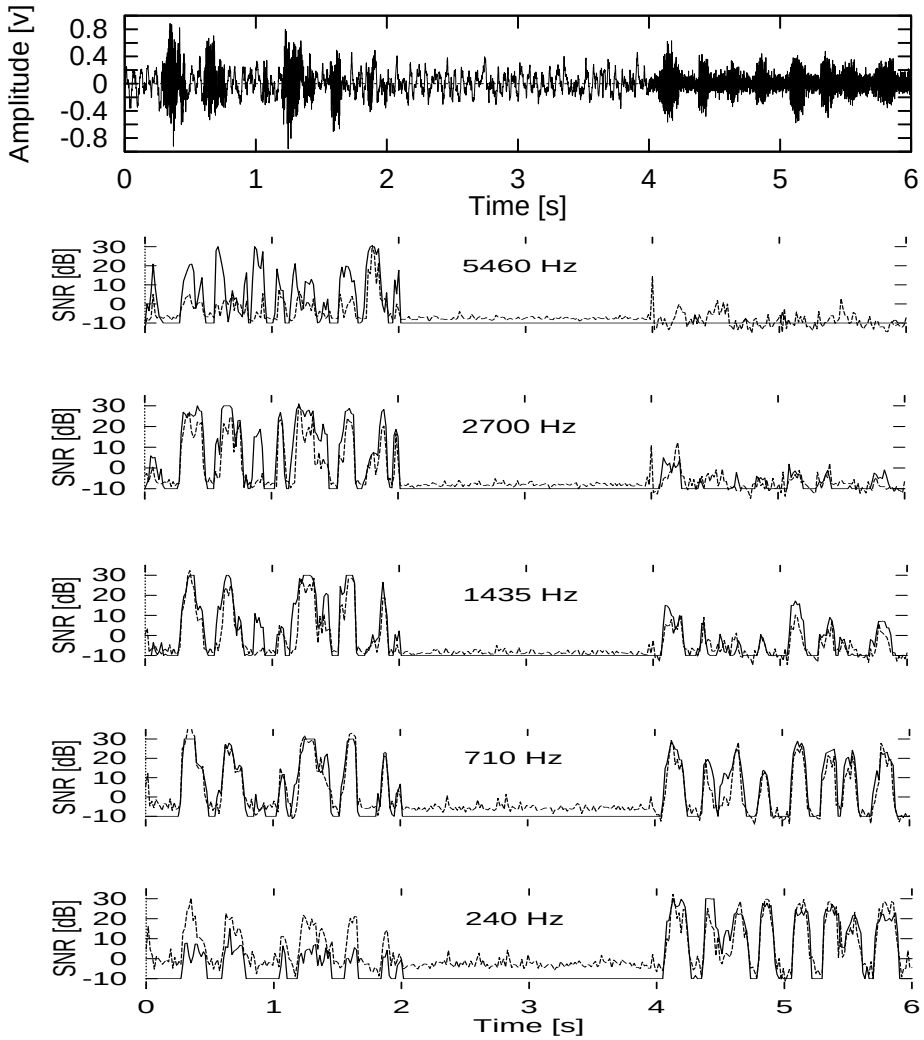

Figure 3: Example for narrow-band SNR estimation. Top: input signal. Below: measured (solid) and estimated (dotted) SNRs as function of time for 5 out of 15 frequency channels. See text.

transformed into 270000 AMS patterns. The long-term SNR between speech and noise during training was 2.5 dB, but the *local* SNR exhibited strong fluctuations (e.g., in speech pauses). The speech material for training was taken from the Phondat database and contained 2110 German sentences from 190 male and 210 female talkers. 41 types of natural noise were taken for training from two different data bases (Alife and Noisex).

For testing, a 36-min mixture of speech (200 speakers, Phondat) and noise (14 types, Alife and Noisex) was taken. These talkers and noise types were not included in the training data.

5. RESULTS

Figure 3 illustrates an example of narrow-band SNR estimation as performed by the above described algorithm. The first panel shows the waveform of the input signal. Between 0-2s, a male talker is speaking in car noise. Between 2-4s, there is a speech pause: only car noise is present. Between 4-6s, the male talker is speaking again, but the noise has suddenly changed from car noise to power drill noise. The following panels show the measured SNR (solid line) and the estimated SNR (dotted line) in 5 out of 15 frequency channels as a function of time. Between 0-2s (speech in car noise), the measured SNR is relatively high in the upper frequency channels due to the low-frequency characteristic of the car noise. Between 2-4s, the measured SNR is at its bottom threshold due to the absence of speech. In the last third, the measured SNR is better in the low-frequency channels due to the high-frequency characteristic of power drill noise. The estimated SNRs (dotted) are in good correlation with the measured ones in drill noise. In car noise, the SNR is over-estimated in low frequency channels, and under-estimated in high frequency channels. The mid-frequency region shows acceptable estimates. The speech pause is detected in all frequency channels.

A more quantitative measure of the accuracy of the algorithm is provided by the mean deviation between the measured SNR and the estimated SNR. The mean deviation was calculated over all AMS patterns generated from the test and training material described in Section 4, for all 15 frequency channels independently. The results are plotted in Fig. 4. The solid line shows the mean deviations for the training material, i.e., it gives an impression on how well the neural net can

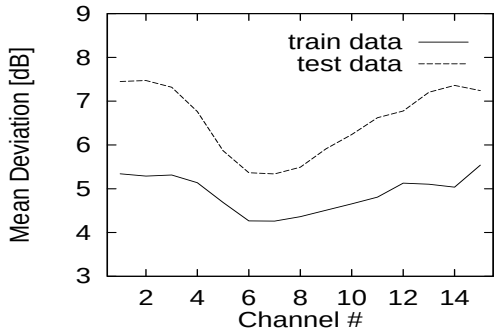

Figure 4: Mean deviation between the measured SNRs and the estimated SNRs as a function of the frequency channel. Calculated over all AMS patterns of the training material (solid) and the test material (dotted) as described in the text.

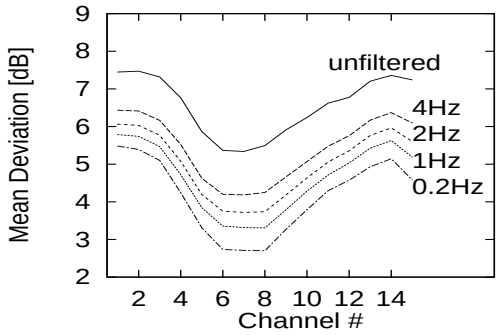

Figure 5: Mean deviation between the measured SNRs and the estimated SNRs for the test material. The temporal evolution of the measured and the estimated SNRs was smoothed with different low pass filters to attenuate fast fluctuations and short-term errors.

reproduce the training data. It can be seen that there still is a distinct discrepancy between targets and outputs after training. The dotted line shows the mean deviations for the „unknown" test data. It can be seen that the estimation accuracy is about 2-3 dB better in the mid-frequency channels, compared to low and high frequency channels. The average deviation between measured SNR and estimated SNR across all frequency channels is around 6.5 dB, which is about 1.5 dB higher than for the training data. This indicates that the network is not „over-trained" and is able to generalize on new data to a certain extend.

For some possible applications of the algorithm, a fast SNR estimation for independent 32ms-frames might not be necessary. Here, temporal smoothing of the estimates can enhance the accuracy of the prediction; „outliers" and short-term errors are attenuated. Figure 5 shows the effect of low pass filtering the temporal evolution of SNR measures and estimates (like those in Fig. 3) prior to calculating the mean deviation. The mean deviation decreases with increasing smoothing, but, of course, the „sluggishness" of the algorithm increases.

6. DISCUSSION

The presented algorithm transforms an incoming sound signal into a series of so-called AMS patterns, which contain both spectral and temporal information about the respective 32ms analysis frames. It could be shown that neural networks can exploit this information to provide an estimate of the present SNR in different frequency channels. There is no *independent* processing of different channels, though, as the network is fully connected and uses information from all channels to provide an SNR estimate for one channel. Future work will concentrate on the extent and importance of this kind of „across-channel" processing. Further improvement of the estimation accuracy might be possible by modifying the network topology. Informal tests showed that increasing the number of hidden neurons decreases estimation errors. However, this requires a larger training set for optimal weight adjusting, which easily leads to problems in terms of computational load during training.

In the future, a noise suppression scheme which bases on narrow-band SNR estimates will be implemented and compared to other noise suppression algorithms in terms of speech intelligibility and ease of listening.

This work was supported by the European Union (TIDE/SPACE).

REFERENCES

(1) Langner, G.: Periodicity coding in the auditory system. *Hear. Res.* **60**, 115-142 (1992)

(2) Kollmeier, B. and Koch, R.: Speech enhancement based on physiological and psychoacoustical models of modulation perception. *J. Acoust. Soc. Am.* **95**, 1593-1602 (1994)

(3) Yang, D., Meyer, G.F., and Ainsworth, W.A.: A neural model for auditory scene analysis. ASA/EAA/DEGA Joint Meeting on Acoustics, Berlin, Germany, 1999, in press

(4) Strube, H.-W. and Wilmers, H.: Noise reduction for speech signals by operations in the modulation spectrum. ASA/EAA/DEGA Joint Meeting on Acoustics, Berlin, Germany, 1999, in press

(5) Tchorz, J. and Kollmeier, B.: Automatic classification of the acoustical situation using amplitude modulation spectrograms. ASA/EAA/DEGA Joint Meeting on Acoustics, Berlin, Germany, 1999, in press