A psychoacoustical model of the auditory periphery as front end for ASR

Jürgen Tchorz and Birger Kollmeier

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

Summary: The application of a psychoacoustical model of the auditory periphery in the field of automatic speech recognition (ASR) is presented. The model was developed to quantitatively predict human performance in typical spectral and temporal masking experiments. Speaker-independent, isolated-digit recognition experiments in different types of noise were carried out to evaluate the robustness of the auditory-based ASR system in adverse conditions. Compared to a standard MFCC front end, the auditory-based preprocessing yielded significantly higher recognition rates in both additive and convolutive noise.

INTRODUCTION

One major problem in the field of ASR is its robustness in noise. Human speech recognition, which - in contrast to ASR - is very robust in noise, is made possible by the interplay between the auditory periphery, which transforms the incoming sound signal into its "internal representation", and the higher auditory processing stages in the brain, which performs the recognition task based on the internal representation. While comparatively little is known about the neural mechanisms of the central auditory processing stages in the brain, much more is known about the peripheral auditory processing stages. Several researchers have proposed algorithms to model different psychoacoustical aspects or physiological processing stages of the auditory periphery. Only few of these models were tested in speech recognition systems, though (1, 2). Compared to standard ASR front ends (MFCC), these auditory-based preprocessing schemes often show only minor advantages in terms of robustness, or they require high computational costs (3).

This paper describes the application of a psychoacoustical model of the auditory periphery as front end for ASR. The model was originally developed to predict human performance in typical spectral and temporal masking experiments (4) but was also applied to different tasks in the field of speech processing (5,6).

PROCESSING STAGES OF THE AUDITORY MODEL

Figure 1 shows the processing steps of the auditory model. The first processing step is a preemphasis of the input signal with a first order high pass filter. This flattens the typical spectral tilt of speech signals and reflects the transfer function of the outer ear. The preemphasized signal is then filtered by a gammatone filterbank using 19 frequency channels equally spaced on the ERB scale with center frequencies ranging from 0.3-4 kHz. The impulse

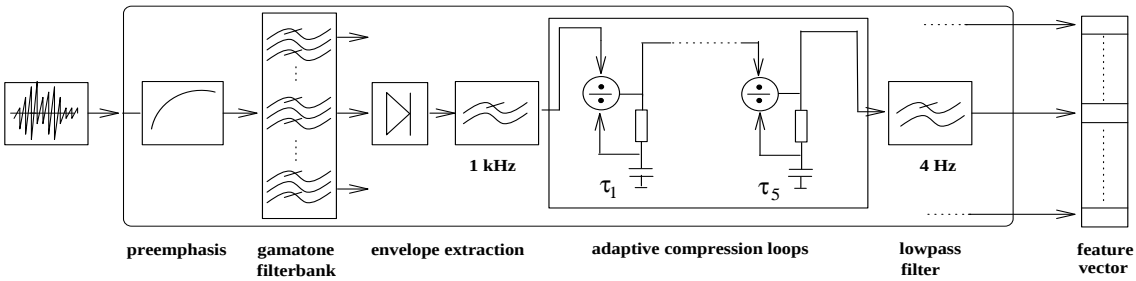

FIGURE 1: Processing stages of the auditory model. See text.

responses of the gammatone filterbank are similar to the impulse responses of the auditory system found in physiological measurements. After gammatone filtering, each frequency channel is halfwave-rectified and first order low pass filtered with a cutoff frequency of 1 kHz for envelope extraction, which reflects the limiting phase-locking for auditory nerve fibers above 1 kHz. Amplitude compression is performed in a following processing step. In contrast to conventional bank-of-filters front ends, the amplitude compression of the auditory model is not static (e.g., logarithmic) but adaptive, which is realized by an adaptation circuit consisting of five consecutive nonlinear adaptation loops. Each of these loops consists of a divider and a RC-low pass filter with an individual time constant ranging from 5-500 ms. Changes in the input signal like onsets and offsets are emphasized, whereas steady-state portions are compressed. Thus, the dynamical structure of the input signal is taken into account over a relatively long period of time. Short term adaptation including enhancement of changes and temporal integration is simulated and allows a quantitative prediction of important temporal effects in auditory perception, such as backward- and forward masking. The step function response of the auditory model is shown in Fig.2.

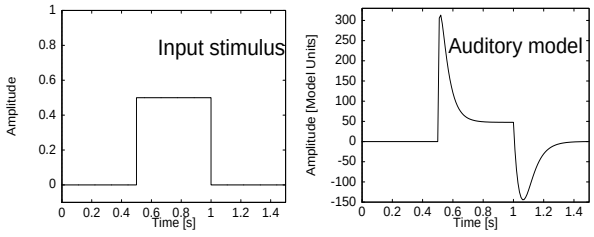

FIGURE 2: Left: input stimulus (envelope). Right: step response of the auditory model (one channel). Initial „overshoot“, transition to compression, and recovery time after stimulus offset.

The last processing step of the auditory model is a first order low pass filter with a cutoff frequency of 4 Hz. It attenuates fast envelope fluctuations of the signal in each frequency channel. Suppression of very slow envelope fluctuations by the adaptation loops and attenuation of fast fluctuations by the low pass filter results in a band pass characteristic of the amplitude modulation transfer function of the auditory model. The maximum amplitude modulation transmission of the model can be found at modulation frequencies around 6 Hz.

RECOGNITION EXPERIMENTS

A number of speaker-independent, isolated digit recognition experiments in different types of additive noise and convolutive distortions were carried out to evaluate the robustness of the auditory-based representation of speech quantitatively.

The speech material for training of the word models and scoring was taken from the ZIFKOM database of Deutsche Telekom AG. Each German digit was spoken once by 200 different speakers (100 males, 100 females). Three different types of noise were added to the speech material at different signal-to-noise ratios before feature extraction: white noise (WN), speech-simulating noise (SN), which was generated from a random superposition of words spoken by a male speaker, and background noise recorded on a construction site (CS). The background noises were scaled and added to the utterances with signal-to-noise ratios of 20, 15, 10, and 5dB. To evaluate the robustness of the auditory front end against convolutive distortions, we applied a band pass filter to the test material which simulates telephone channel conditions in a further run. This stationary convolution in the time domain is a special case of convolutive noise encountered in most telecommunication applications. The filter was realized using a combination of a high pass filter with a cutoff frequency of 330 Hz and a low pass filter with a cuttoff frequency of 3300 Hz. Both filters were fourth-order Butterworth filters. For training and testing, we used a standard continuous-density HMM recognizer with 5 Gaussian mixtures per state, diagonal covariance matrices and 6 emitting states per word model. The word models were trained with features from 100 undisturbed utterances of each digit. Features for testing were calculated from another 100 utterances of each digit which were distorted by additive or convolutive noise before preprocessing. As control front, end we used mel frequency cepstral coefficients, which are widely used in common ASR systems. The coefficients were calculated from Hamming-windowed, preemphasized 32ms segments of the input signal with a frame period of 10ms. In our experiments, each mel cepstrum feature vector contained 26 features (12 coefficients, log energy, and the respective first temporal derivatives).

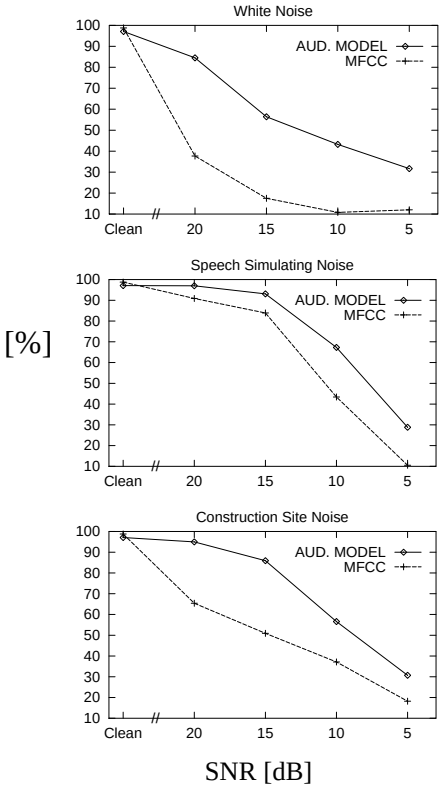

FIGURE 3: Recognition rates in % in different types of additive noise yielded with the auditory model and with the control front end (mel-scale cepstral coefficients, MFCC) as function of signal-to-noise ratio in dB.

RESULTS

The speaker-independent digit recognition rates in clean speech and in additive noise obtained with the auditory preprocessing and the control front end are shown in Fig. 3. The recognition rates in per cent are plotted as a function of the signal-to-noise ratio in dB. In undisturbed speech, the control front end yields a higher recognition rate (98.8 %) than the auditory-based front end (97.1 %). In additive noise, however, the auditory features are more robust than those of the control front end. Even in only slightly disturbed speech (20 dB SNR), the recognition rates obtained with the auditory model are significantly higher in all tested types of noise.

The digit recognition rates in convolutive distortions are shown in the second row of Table I. It can be seen that the representation of speech from the auditory-based front end is only gradually affected by simulating telephone channel conditions, compared to the control front end.

TABLE 1: Recognition rates [%] in convolutive noise yielded with the auditory model and the control front end.

	Aud. model	**MFCC**
Clean speech	97.1	98.8
Convolutive noise	92.2	68.5

DISCUSSION AND CONCLUSION

The psychoacoustically-motivated auditory model which was originally developed to describe human performance in typical psychoacoustical spectral and temporal masking experiments allows more robust digit recognition rates in noise, compared to a standard MFCC front end. In the auditory model, the enhancement of onsets and offsets, and the compression of steady-state portions by the adaptation loops play a major role in the computation of robust features (7). The performance in noise can further be enhanced by using locally recurrent neural networks as pattern recognition stage instead of HMM recognizers (8), as well as applying explicit noise suppression algorithms (e.g., spectral subtraction) prior to feature extraction (9).

However, to further evaluate the potential of the auditory model in speech recognition systems, experiments with large word vocabulary as well as connected word recognition experiments are necessary.

REFERENCES

(1) Ghitza, O., *J. Phonetics* **16**, pp. 109-123 (1988).
(2) Seneff, S., *J. Phonetics* **16**, pp. 55-76 (1988).
(3) Jankowski, C.R., *IEEE Trans. Speech and Audio Processing* **3**, pp. 286-293 (1995).
(4) Dau, T., Püschel, D., and Kohlrausch, A., *J. Acoust. Soc. Am.* **99**, pp. 3615-3622; pp. 3623-3631 (1996).
(5) Hansen, M. and Kollmeier, B., „Using a Quantitative Psychoacoustical Signal Representation for Objective Speech Quality Measurement," in: *Proc. Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, Munich, Germany, pp. 1387-1391, 1997.
(6) Holube, I. and Kollmeier, B., *J. Acoust. Soc. Am.* **100**, pp. 1703-1716 (1996).
(7) Tchorz, J. and Kollmeier, B., *J. Acoust. Soc. Am.* (submitted) (1999).
(8) Tchorz, J., Kasper, K., Reininger, H., and Kollmeier, B., „On the Interplay Between Auditory-Based Features and Locally Recurrent Neural Networks," in*: Proc. EUROSPEECH '97*, Rhodes, Greece, pp. 2075-2078, 1997.
(9) Kleinschmidt, M., Wittkop, T., and Kollmeier, B., „Evaluation of monaural and binaural speech enhancement for robust auditory-based automatic speech recognition," *Joint Meeting ASA/EAA/DEGA*, Berlin, Germany, 1999.