Auditory principles in speech processing – do computers need silicon ears ?

Birger Kollmeier

Medizinische Physik, Universität Oldenburg, D-26111 Oldenburg, Germany,
http://medi.uni-oldenburg.de
birger.kollmeier@uni-oldenburg.de

Abstract

A brief review is given about speech processing techniques that are based on auditory models with an emphasis on applications of the "Oldenburg perception model", i.e., objective assessment of subjective sound quality for speech and audio codecs, automatic speech recognition, SNR estimation, and hearing aids.

1. Introduction

The human auditory system has solved a number of problems that speech communication engineers are still struggling with, such as: localizing sound sources and suppressing reverberation in an arbitrary environment, judging the quality of lossy speech and audio transmission channels without knowing the source signal and recognizing speech under adverse acoustical conditions. It thus seems advisable to exploit auditory processing principles known from models of hearing and to implement them into speech processing applications.

This contribution focuses on the most relevant auditory processing principles and reviews their application to speech quality prediction, noise reduction for hearing instruments, and robust automatic speech recognition. To characterize the "effective" signal processing performed by our ear, a functional model is employed which was originally designed to predict a variety of psychoacoustical effects. Its application, e.g., as a front end for robust speech recognition or as a tool for audio and speech coding quality assessment is introduced and discussed. By following the ear´s construction principles, the performance and robustness of speech communication systems can considerably be enhanced.

2. Models of the "effective" signal processing in the auditory system

Rather than analyzing each aspect of neural signal processing in the auditory system with sophisticated models of neural assemblies, models of the "effective" signal processing in the auditory system concentrate on the overall performance of the auditory system. Hence, the transformation of the acoustic input signal into its "internal representation" can be described using linear or nonlinear signal processing elements that are motivated by physiological and psychoacoustical data. Several models of this kind have been proposed in the literature (i.e. Munich loudness model [25], Boston model [2], Cambridge model [19]). In the following we concentrate on the "Oldenburg perception model" [3], [4] which puts a special emphasis on amplitude modulation processing and the modulation filterbank concept, i.e. the analysis of modulation frequencies for the envelope within each critical band [14]. The internal representation computed by the model is supposed to be used by our cognitive system (modeled by an "optimal detector") to detect the appropriate stimulus. A sketch of such an auditory model is given in Fig. 1.

The action of the cochlea, i.e. the separation of different frequencies in a series of "critical band" filters is represented by a bank of bandpass filters. The inner and outer hair cells perform an envelope extraction. It is followed by dynamic compression, i.e. an amplification at low levels and less amplification or even attenuation at high levels. Parts of this dynamic compression is performed by the outer hair cells in a very fast way, but also by subsequent stages of the auditory system contribute with larger time constants. Subsequently, the information from both ears is combined by performing a binaural comparison between both sides. The accuracy of the whole processing is limited by imperfections of the neural representation of the stimuli. This is modeled by the addition of "internal" noise before the complete pattern (from both monaural systems

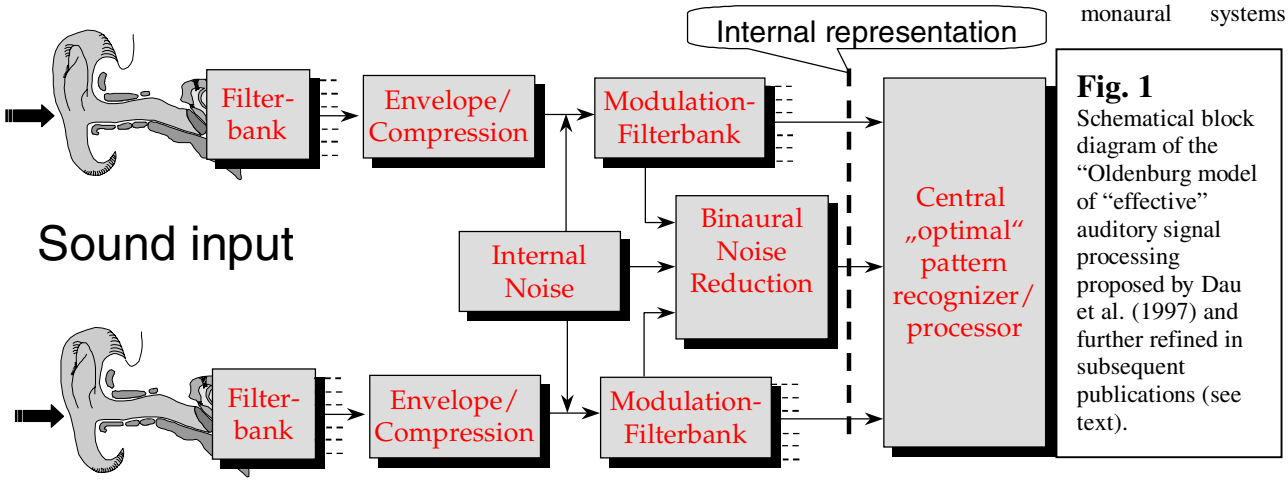

Fig. 1 Schematical block diagram of the "Oldenburg model of "effective" auditory signal processing proposed by Dau et al. (1997) and further refined in subsequent publications (see text).

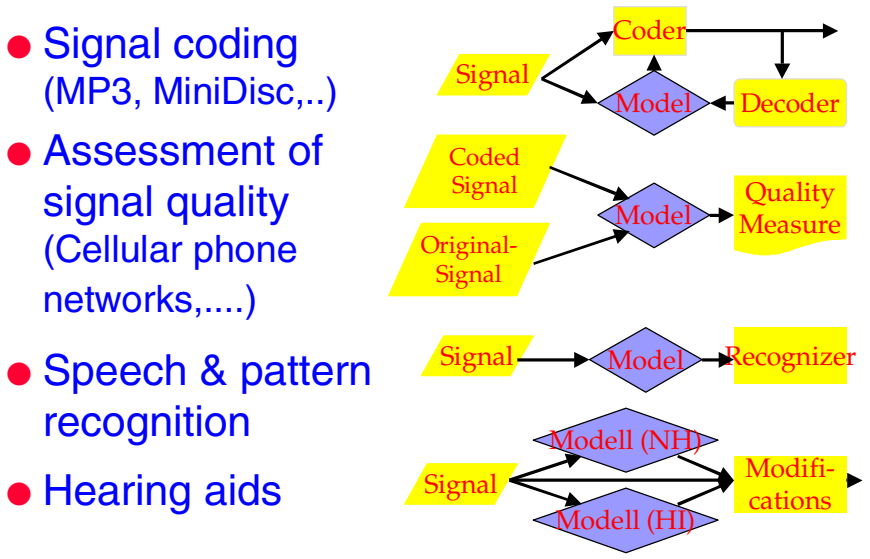

Fig. 2: Applications of auditory models to speech pocessing (schematic).

and from the binaural stage) reaches the central pattern recognizer which is modeled as an "optimal" detector. The model has been validated by its ability to predict a wide range of psychoacoustical and speech perception tasks both by normal and hearing-impaired listeners [3],[4].

Application of the model

The knowledge about the processing principles in audition can be directly applied to technical systems that deal primarily with speech processing, i.e., automatic speech recognition, noise suppression, and digital hearing aids. Fig. 2 gives a schematical view how this can be achieved for several applications. The most widely accepted application of auditory models is speech coding and speech synthesis (first introduced by Schroeder & Atal, c.f. [20]) and audio coding [1] where a comparatively simple auditory model is used to assess the perceivable difference between the original signal and the coded/decoded signal. Hence, the codec is constructed to produce a mimimum difference at the output of the auditory model.

A similar setup with an auditory model (especially the "Oldenburg Perception model") has successfully been used to objectively predict the subjective quality of nonlinearly distorted speech (such as mobile phone communication, [5]), the subjective quality of hearing aid processing schemes [15],[16], and the ability of audio codecs to preserve the acoustical quality of music [10]. The respective quality measure is derived by comparing the representation of the original acoustical signal with the one of the nonlinearly processed signal at the *output* of the auditory model rather than using features of the acoustical signas themselves.

Furthermore, auditory processing principles such as the modulation filterbank concept have been shown to be advantageous in (monaural) noise reduction techniques, i.e., by directly estimating the respective speech-to-noise from the amplitude modulations spectrogram [13], [22], and subsequently using this estimate to reduce noise [23]. Similarly, principles of binaural processing (i.e., comparison between left and right ear) have been shown to be advantageous in (binaural) noise reduction techniques for "difficult" acoustical situations when two signal channels are available (such as, e.g. bilateral hearing aids or stereo recordings) [24],[18]. The importance of basic auditory-model based signal processing led to a cooperation with the Computer Science faculty with the aim to implement the Oldenburg Perception model in hardware in a power-saving way ("Silicon Ear").

In addition, the "Oldenburg Perception model" has successfully been used as a front end for an automatic speech recognizer to enhance speech recognition under noisy conditions [22], [11]: The auditory-model-based "internal representation" of (noisy) speech signals at the output of the model is used as the input pattern to a speech recognizer, i.e., a pattern recognizer using for example a neural network classifier.
The main reason for this approach is that the technical representation of speech currently used in the computer is not appropriate even for simple speech recognition tasks that are performed quite easily by our brains. The idea therefore is to use in the computer the same representation of acoustical signals and speech as in our brain. Similar arguments have already led to successful concepts to improve the robustness of speech recognizers for real-life recordings, such as RASTA processing [7],[6]. Even more features of state-of-the-art speech recognizers can also be found in the auditory system, such as, e.g., the logarithmic representation of frequencies similar to the Bark or Mel scale, and logarithmic compression and online normalization similar to loudness perception. These features have originally been introduced for technical reasons and not necessarily to copy the human ear by a machine speech recognizer.

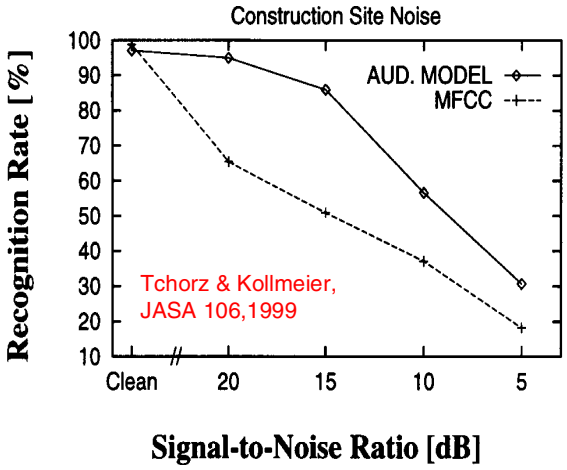

Fig. 3: Speaker-independent digit recognition rate in % in construction site noise obtained with an auditory model front end and with the control front end (mel-scale cepstral coefficients, MFCC) as a function of signal-to-noise ratio in dB (From [21]).

To demonstrate the advantage of an auditory-model driven approach in our case, an auditory front end in combination with a standard HMM speech recognizer was compared to the standard front end using a mel-frequency-cepstral-coefficient (MFCC) representation. In quiet (i.e., for clean speech) no substantial difference in recognition rate occurs. However, the performance of the conventional speech recognizer drops down significantly if the input signal is corrupted by noise. In the auditory model version, however, the performance stays up and finally degrades at a more unfavorable signal-to-noise ratio than the conventional system.

The fourth application of auditory models depicted in Fig. 2 are "intelligent" digital hearing aids: The idea is to compare the output of the model for a normal listener with the output for a hearing impaired listener and to derive those modification that should be introduced in order for the hearing-impaired listener to perceive approximately the same auditory image as the normal listener does. A more systematic treatment of this "model-based hearing aid" is given by [9].

3. Conclusion: Which auditory features are necessary?

Even though a copy of the "effective" signal processing in the auditory system appears to be advantageous for a number of speech processing applications (as outlined above), this still does not mean that each aspect of the auditory system should be integrated into futural speech processing systems. Rather, it appears that certain features of auditory processing should be adopted in technical systems that are most relevant for the extraordinary good and robust performance of the human auditory system in acoustically "difficult" situations. Candidates for these "prime auditory processing features" are:

- constant relative bandwith (i.e., self-similar) filterbank design: Similar to third-octave filterbanks that are common in acoustic analysis, auditory – processing based filterbanks show a constant relative bandwidth at high frequencies and a more or less constant absolute bandwith at low frequencies (as reflected by the Bark- or the mel-scale or the more recent ERB-scale [17]). An implementation for an auditory filterbank well-suited for practical purposes has been provided by [8].
- Adaptive dynamic compression: For steady-state signals the (quasi-) logarithmic compression of acoustic signal energy in the auditory system is already represented by a logarithmic transform in many technical systems. However, a time-dependent adaptation mechanism performs an adaptive, time-variant compression in the auditory system which leads to pre- and postmasking effects and has been represented in the "Oldenburg perception model" by a series of nonlinear adaptation loops. In technical system, such an adaptive compression can be approximated by a modulation bandpass filter (as used in RASTA processing [7] or by appropriate online normalization schemes.
- Modulation spectrogram: Spectral decomposition of amplitude modulations in each frequency band is employed by our auditory system as an elegant way to encode the temporal (envelope) characteristics of a stimulus into a "spatial" feature. Moreover, the synchrony of amplitude modulations in different frequency band is used by the auditory system to "bind" together the spectrally separated components of an acoustical object. In technical system, the benefit of using the amplitude modulation spectrogram as an input to a pattern recognizer for automatically recognizing acoustical objects or scenes has still to be demonstrated in a broader range of applications than covered in this contribution.
- Second order time-frequency patterns: Recent investigations [12] indicate that using specific spectro-temporal prototype patterns (such as Gabor functions) provide more robustness for automatic speech recognizers. Interestingly, such patterns seem to be correlated to perceptual "elementary units" or second-order features used by the auditory system.
- Binaural noise reduction: Our head appears to utilize the simultaneous acoustic inputs to both ears like an optimal two-sensor-adaptive beamformer that can cancel out one spatial direction per unit of time. While little is still known how the brain actually steers this beamformer, some technical systems have already devised similar mechanisms to perform a noise reduction based on stereo input signals ("Cocktail party processing") [18],[24].

In conclusion, even though the good performance of auditory-model based speech processing approaches would suggest that computers need "silicon ears", it still may be sufficient to implement only the most relevant aspects of the "effective" auditory signal processing in technical systems in order to obtain improved performance in speech processing. New insights into our ears will therefore necessiate new speech processing algorithms.

4. Acknowledgements

Supported by Deutsche Forschungsgemeinschaft (DFG). The author thanks all members of the Medical Physics group for their support and help.

5. References

[1] Brandenburg, K. and G. Stoll, ISO-MPEG-1 Audio: A Generic Standard for Coding of High-Quality Digital Audio. *J. Audio Eng. Soc,* 42: 780-792, 1994.

[2] Colburn, H. S. in: Auditory Computation (Springer, New York, 1996), p. 332.

[3] Dau, T., B. Kollmeier, and A. Kohlrausch, Modeling auditory processing of amplitude modulation: II. Spectral and temporal integration. *J. Acoustical Soc. Am.,* 102(5): 2906-2919, 1997.

[4] Derleth, R.-P., T. Dau, and B. Kollmeier, Modeling temporal and compressive properties of the normal and impaired auditory system. *Hearing Research,* 159(1-2): 132-149, 2001.

[5] Hansen, M. and B. Kollmeier "Objective modeling of speech quality with a psychoacoustically validated auditory model." *J. Audio Eng. Soc.* 48(5): 395-408,2000.

[6] H. Hermansky, "Should recognizers have ears?," *Speech Communication*, vol. 25, pp. 3–24, 1998.

[7] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Trans. SAP*, vol. 2, no. 4, pp. 578–589, 1994.

[8] Hohmann, V., Frequency analysis and synthesis using a Gammatone filterbank. *Acustica / acta acustica*, 88(3):, p. 433-442. 2002

[9] Hohmann, V., Modellbasierte Signalverarbeitung in heutigen und zukünftigen Hörgeräten. *Zeitschrift für Audiologie/Audiological Acoustics,* Suppl. IV: 136-140, 2001.

[10] Huber, R. and B. Kollmeier, *Vorhersage von Audioqualität mit einem psychoakustischen Modell*, in *Fortschritte der Akustik - DAGA 2001*. 2001, DEGA e.V.: Oldenburg. p. 468-469.

[11] Kleinschmidt, M., J. Tchorz and B. Kollmeier "Combining Speech Enhancement and Auditory Feature Extraction for Robust Speech Recognition." *Speech Communication* 34(1-2: Special Issue on Robust ASR): 75-91, 2001

[12] Kleinschmidt, M. and D. Gelbart. *Improving Word Accuracy with Gabor Feature Extraction*. in *ICSLP 2002*. Denver

[13] Kleinschmidt, M. and V. Hohmann, Sub-band SNR estimation using auditory feature processing. *Speech Communication,* 39(1-2 (Special Issue on Speech Processing for Hearing Aids)): 47-64, 2003.

[14] Kollmeier, B. and R. Koch, Speech Enhancement Based on Physiological and Psychoacoustical Models of Modulation Perception and Binaural Interaction. *J. Acoust. Soc. Am.,* 95: 1593-1602, 1994.

[15] Marzinzik, M. and B. Kollmeier, Speech pause detection for noise spectrum estimation by tracking power envelope dynamics. *IEEE Transactions on Speech and Audio Processing,* 10(2): 109-118, 2002.

[16] Marzinzik, M. and B. Kollmeier, Predicting the Subjective Quality of Noise Reduction Algorithms for Hearing Aids. *Acta acustica/Acustica,* 89: 521-529, 2003.

[17] Moore, B.C.J., *An Introduction to the Psychology of Hearing*. fourth ed. 1997: Academic Press.

[18] Nix, J., Hohmann, V. (submitted). "Statistics of interaural parameters in real sound fields employing one directional sound source and its application to sound source localization." J. Acoust. Soc. Am.

[19] Patterson, R.D., A.M. H., and C. Giguère, Time-domain modeling of peripheral audotory processing: A modular architecture and a software platform. *J. Acoust. Soc. Am.,* 98(4): 1890--1894, 1995.

[20] Schroeder, M.R., *Computer Speech*. 1999, Berlin: Springer.

[21] Tchorz, J. and B. Kollmeier "A model of auditory perception as front end for automatic speech recognition." Journal of the Acoustical Society of America 106(4): 2040-2050. 1999

[22] Tchorz, J. and B. Kollmeier, " Estimation of signal-to-noise ratio with amplitude modulation spectrograms. *Speech communication* 38: p. 1 – 17,2002

[23] Tchorz, J. and B. Kollmeier (in press). "Noise suppression based on amplitude modulation analysis." *IEEE Transactions on Speech and Audio Processing.*

[24] Wittkop, T., Hohmann, V., Strategy-selective noise reduction for binaural digital hearing aids. *Speech Communication,* 39: 111-138, 2003.

[25] Zwicker, E. and H. Fastl, *Psychoacoustics : Facts and Models*. 2 ed. 1999, Berlin: Springer.