Acoustics'08
Paris
June 29-July 4, 2008
www.acoustics08-paris.org

Improving speech intelligibility based on a conjunction of multiple perceptual models

Anton Schlesinger and Marinus Boone

University of Technology Delft, Lorentzweg 1, 2628 CJ Delft, Netherlands
m.m.boone@tudelft.nl

A method for the improvement of speech intelligibility by suppression of lateral noise sources and reverberation is presented. The method is based on computational auditory scene analysis. Lateral inhibition is achieved by applying a psychoacoustical model of binaural interaction and modulation perception. Dereverberation is performed by identifying and emphasizing time- and frequency-regions that belong to the direct sound as a function of interaural coherence. The processing scheme is assessed with modulation transfer metrics in adverse acoustical situations. In a preliminary implementation, a small but constant gain in speech intelligibility is observed.

1 Introduction

Considering the success of hearing aids, the enhancement of speech intelligibility (SI) is a crucial issue. The enhancement of SI is essentially dependent on the enhancement of the signal-to-noise ratio (SNR) at the input of the auditory system. The success in the process is bound to the maintenance of speech quality (also called listening-comfort). A balance between these two objectives by signal processing can be achieved by computational auditory scene analysis (CASA). In this regard, functional approaches that mimic the mammalian auditory processing in its ability to characterize and separate the sound scape while obeying the continuity of signal processing are applied [12].

In this work a linkage of auditory mechanisms that are attributed to enhance the SI in poor acoustical situations is presented. At first we employ a model of binaural interaction and modulation perception to efficiently inhibit lateral noise sources. It has been shown, that exploiting modulation in connection with binaural cues in speech signals offers a robust means to dismantle an acoustic scene in the presence of reverberation and noise [6]. A second mechanism that is incorporated here, renders aspects of the precedence effect to suppress echoes. A dereverberation-stage is implemented by calculating the interaural coherence (IC) that enables the separation of free-field cues from diffuse-field cues. By subsequently weighting monaural cues, direct sounds attain precedence over echoes. The mechanisms where combined in a sequential mode and assessed with the modulation transfer metrics. Section 2 outlines the design of the algorithm and section 3 gives results of the assessment of the proposed processing scheme.

2 Background

The design of a hearing aid is subject to its computational load and its applicability in real-world situations. When employing CASA these conditions implicate a bottom-up approach, which corresponds to the natural abstraction process of the auditory scene analysis.

A bottom-up auditory scene analysis is generally understood as a two-stage process. In a first stage the sound scape is decomposed into auditory events (also denoted auditory objects). The second stage groups auditory objects that emanate from the same environmental event (also known as the *binding problem*). In order to perform this scene analysis, the brain-stem analyses in four nuclei different cues that are inherent to auditory objects [12]. By analyzing perceptual cues as pitch, beating and interaural differences, a three-dimensional neural display images a tonotopical, a periodotopical and a topological representation of the auditory space [6]. By intentional selection of the cerebrum an arbitrary auditory object of interest is subsequently forwarded to the speech areas. Alongside with the buildup of the orthogonal display to bind an auditory scene, each nuclei performs a perceptual weighting with respect to direct sound and reflections. Accordingly, perceptual cues that belong to the direct sound are equipped with a higher perceptual weight than cues that belong to reflections. Among other phenomenons, this neural processing is attributed to the precedence effect [7].

An algorithm for the enhancement of SI in speech signals that utilizes psychoacoustical models of binaural interaction and modulation perception was successfully implemented by Kollmeier and Koch (KK) [6, 13] and forms here the basis for the presented combined processing scheme. The KK-algorithm yields a constant gain in SNR even in poor acoustical situations without introducing artifacts and is therefore attractive for an integration in hearing aids.

Regarding an implementation of aspects of the precedence effect for the enhancement of the SI, several combined processing schemes that included aspects of echo-suppression [11, 1, 13] were developed. Yet, the impact of only the echo-suppression to the improvement of SI has not been fully assessed. This led to the motivation to implement different models of the precedence-effect and to study their effects on SI, individually.

To functionally describe the precedence-effect in its diversity, a set of progressional models was developed. For the application of aspects of the precedence-effect in hearing aids, we examined three models of different complexity.

2.1 Review of different approaches

The first model of precedence effect that was tested here, is the computational model for echo suppression by Martin [8]. For each centre-frequency, the model generates an onset-mediated inhibition, which lasts (triggered by a transient pulse) approximately 10 ms. The model is especially appealing due to its ease of application. In our implementation we incorporated an inner hair cell model by Meddis [9] and used an implementation of it by Wang and Brown [12]. We reduced the spontaneous firing rate and therewith a constant inhibition according to the parameters in [9] and lowered the inhibition in order to apply to the principles of continuity preserving signal-processing.

The processing was assessed with 2 min of speech in different reverberant conditions (0.3 s to 1.33 s) in a simulation of a small room [10] as depicted in Fig. 2 (without interference). No improvement in SI, as assessed with the modulation transfer metrics, was observed. This is in accordance with the general knowledge about the SI supporting influence of early reflexions in the presence of late reflections[1].

The second and third model that were applied here to enhance SI resemble each other in their implementation. The only difference is the application of an inner hair cell model after the peripheral processing for the model of Faller and Merimaa (FM) [3], which is missing in the implementation of Peissig [5, 13]. After a peripheral processing by a gammatone-filterbank, the model of Peissig and the adapted model of FM determine by a normalized cross correlation among left and right channel for each centre-frequency the interaural coherence (IC), interaural level difference (ILD) and interaural time difference (ITD) as described in [3]. The IC is used to define free-field cues and therewith to weight the monaural centre-frequency and time representation. ILD and ITD cues are verified with respect to reference values, in order to additionally suppress lateral noise sources. By a weighting-function that uses the triplet IC, ILD, ITD as an input, the algorithms can be tuned to different suppression characteristics between dereverberation and echo-suppression [5, 13].

Analogue to the assessment of the computational model of Martin, we tested the FM- and Peissig-model with pure speech in different reverberant conditions. When adjusting the algorithms to only perform a dereverberation, no overall improvement of the SI was observed. When utilizing ILD and ITD cues to perform a binaural beamforming, a small improvement in SI was observed for the FM-model and a more pronounced enhancement of SI was recorded for the Peissig-model. More studies have to be performed to assess the first results of the processing correctly. To give a first explanation, we think that the dereveberation stage should undergo an expansion of the weighting-function in order to act stronger. Moreover, it is known from literature [12] that moderate reverberation does not deteriorate SI of a single source. As the FM-model enhances onsets and adapts to constant sounds, ILD and ITD are less beneficial cues for lateral suppression. Motivated by the success of the Peissig-model, it became the candidate for a connection of echo-suppression with the elaborate KK-model of binaural interaction and modulation perception.

2.2 Design of the algorithm

The proposed algorithm is depicted in Fig. 1. Defining an arrangement and supported by the fact that modulation perception is relatively robust in reverberation (i.e., a limited amount of reverberation does not deteriorate speaker identification), the KK-model was chosen as a first stage to perform a robust lateral suppression of noise sources and interfering speakers.

After the peripheral processing with a 4th order gammatone-filterbank (denoted cochlear-domain in Fig. 1), the envelope is extracted via the Hilbert-Transform. By a Fourier-Transform along time-frames, an orthogonal display of centre-frequencies and modulation-frequencies is established (denoted modulation-domain in Fig. 1). Using the modulation-domain of the left and right channel, interaural time- and level differences are calculated and compared with reference-values to maintain only auditory objects that emanate in the line of vision. The original modulation-domain is then multiplied with the weighting-factors as determined by the binaural beamformer. After an inverse Fourier-Transform, the original band-passed signal in the cochlear-domain is modulated with an altered envelope.

The output of the KK-model is then fed into the model of Peissig. From literature [13] it is known that the Peissig-algorithm performs better at favorable SNR. In such a way the overall achievement in SNR is increased, in addition to the inherent lateral- and echo-suppression. In a first implementation of the binaural filtering we employed equal trapezoidal weighting-functions with equal ITD_{ref} (± 0.3 ms) and ILD_{ref} (± 1 dB) for calculating the weighting factors 1 in Fig. 1 and weighting factors 2 as outlined in [6]. Reliable binaural cues were identified with the standard deviation for calculating weighting factors 1 in Fig. 1 as outlined in [6]. The weighting factors 2 are calculated from the triplet IC, ILD and ITD. Either weighting functions were finally expanded by squaring.

3 Simulation and Results

To assess the processing of the proposed processing scheme of suppression of lateral noise sources and reverberation a setup as shown in Fig. 2 was simulated in a virtual environment using head-related transfer functions of a manikin [10]. A small room was chosen to analyze the suppression of early reflections. The manikin was heading a target speaker and at −90°, a noise source with the long term speech spectrum of the target speaker was placed.

In order to determine the influence of the proposed algorithm to the SI, modulation transfer metrics (M) were calculated with an envelope regression method [4] in seven octave-bands with centre-frequencies ranging from 125 Hz to 8 kHz and transformed into the apparent SNR ($aSNR$):

$$aSNR = 10log_{10}\left(\frac{M}{1-M}\right). \tag{1}$$

As a probe stimulus, 2 min of speech (spelt alphabet), spoken by a male talker from the ShATR corpus [2] were recorded and processed by the proposed algorithm. The sampling frequency was 16 kHz and the peripheral processing was performed with a gammatone-filterbank with 35 bands. The results are shown in Figures 3 for RT = 0.5 s and RT = 1.33 s in Fig. 4. The results refer to an improvement of the SI by the combined model of auditory mechanisms at an $SNR = -4$ dB. The progression of the curves exposes a maximum at the octave-band with a centre-frequency of 250 Hz. In this frequency-region resides the first formant of speech sounds, which usually carries a compaction of energy and thus lowers the local SNR. An explanation of the local minima for the processing of the KK-algorithm

[1] Cf. with the room acoustical measure of SI (C_{50}: *Deutlichkeitsmass*) that assigns all energy that follows an impulse in a time-frame of 50 ms as supporting and later arriving energy as detrimental to SI.

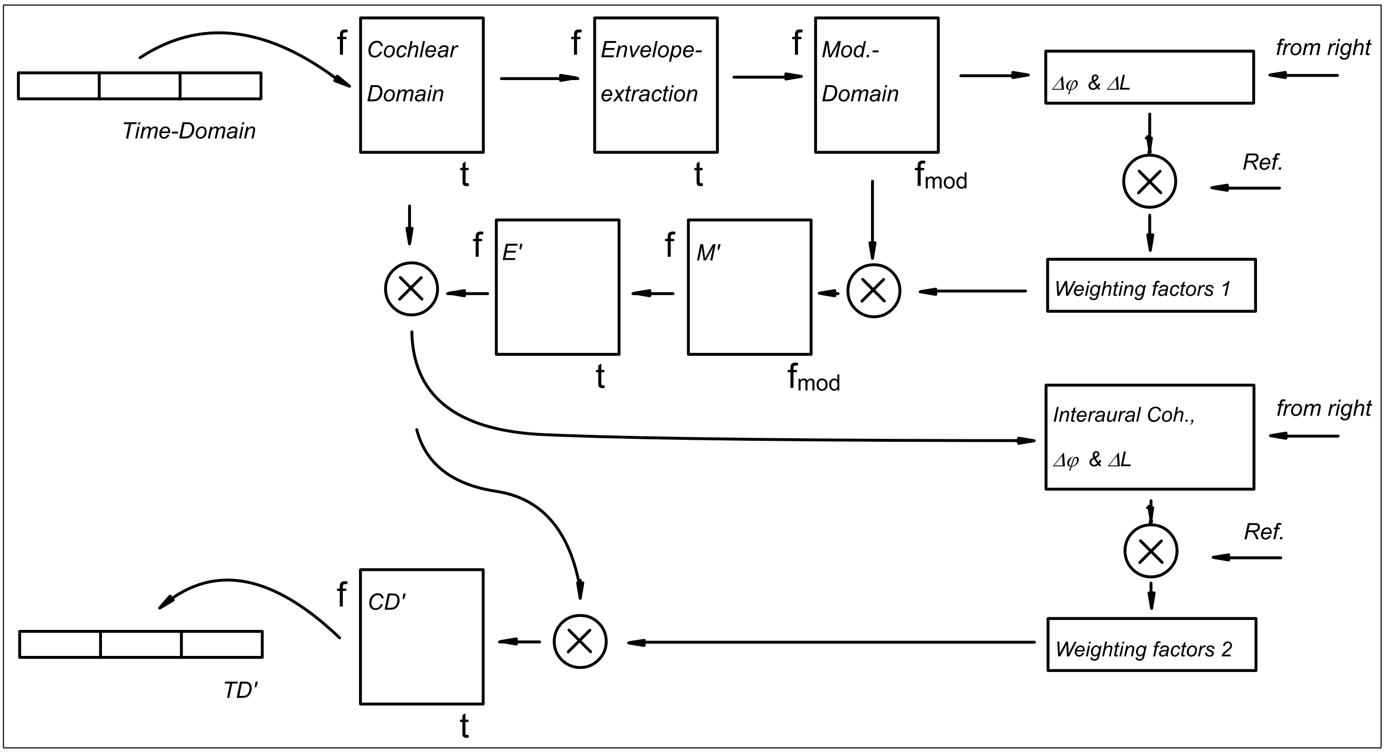

Figure 1: Left channel of the combination of the Kollmeier-and Koch-algorithm [6] (upper half) and the Peissig-algorithm [13] (lower half) to suppress lateral sounds and reverberation.

has to be left to further investigation. The overall observed increase in SNR is smaller than expected from literature [6, 13]. This might be a consequence of the preliminary state of the implementation and its loosely chosen reference-parameter of the binaural beamforming. Nevertheless the increase in SNR is constant in a range of poor acoustical situations with different SNR that are not reproduced here for brevity and therefore the here presented preliminary implementation of a combined model shows its potential in enhancing SI. By applying an iterative adjustment process including the vast amount of parameters the gain in SNR should significantly increase.

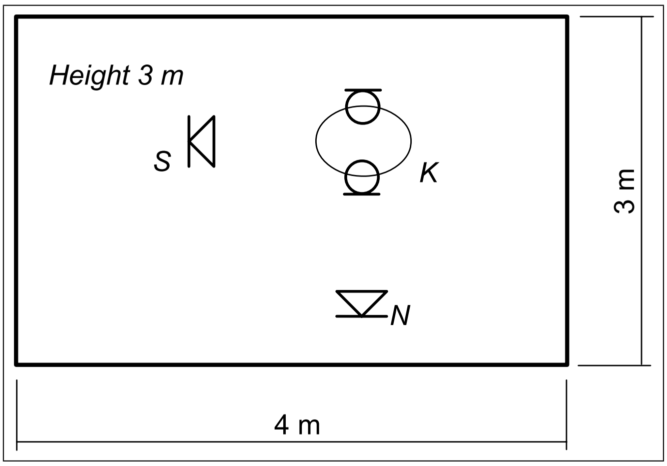

Figure 2: Schematic sketch of simulated setup, S denotes the target-speaker, N the noise-signal and K the head.

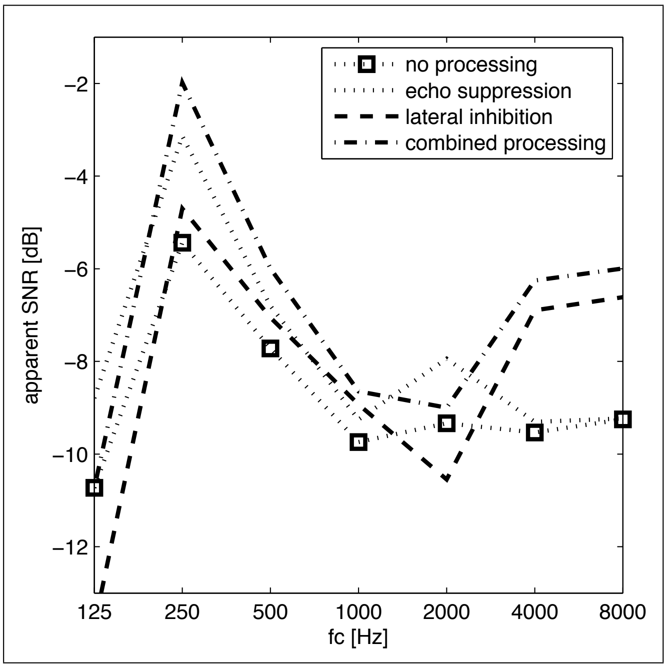

Figure 3: Apparent SNR for 7 octave-bands at the left ear for different processing stages, at $RT = 0.5$ s, lateral suppression denotes the KK-algorithm, echo-suppression denotes the Peissig-algorithm.

4 Conclusion

In this work a CASA processing scheme which is based on models of binaural interaction, modulation perception and aspects of the precedence effect is presented. In a tentative combination of an algorithm that incorporates those models, an increase in SI was observed by assessing the algorithm with the modulation trans-

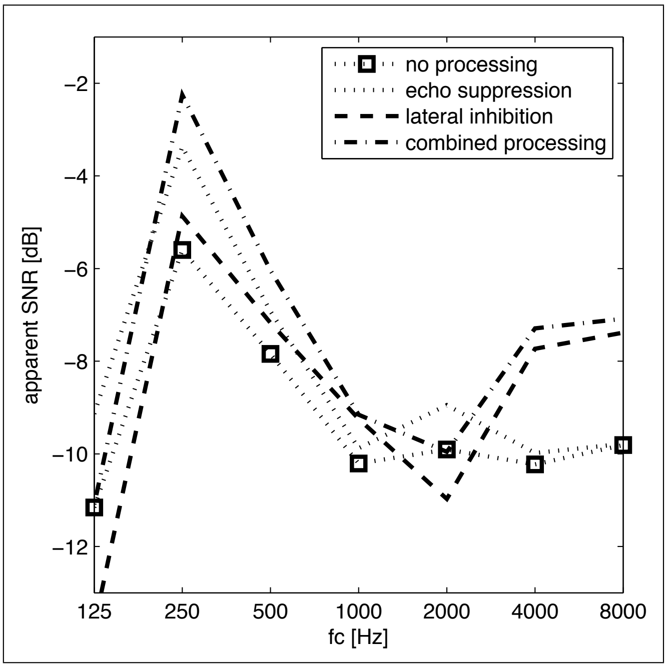

Figure 4: Apparent SNR for 7 octave-bands at the left ear for different processing stages, at $RT = 1.33$ s, lateral suppression denotes the KK-algorithm, echo-suppression denotes the Peissig-algorithm.

fer metrics. The increase in SI is low but constant and therewith the combined processing scheme of several auditory mechanisms forms a starting point for further investigation regarding a future application in hearing aids.

References

[1] Albani, S. et al.: Model of binaural localization resolving multiple sources and spatial ambiguities., In: Psychoacoustics, speech and hearing aids., *World Scientific*, 1996.

[2] Crawford, M. D.: Design, collection and analysis of a multi-simultaneous-speaker corpus., *Proc. Inst. Acoustics*, 1994.

[3] Faller, C. and Merimaa, J.: Sound localization in complex listening situations: Selection of binaural cues based on interaural coherence., *J. Acoust. Soc. Am.*, 2004.

[4] Goldsworthy, R. L. and Greenberg, J. E.: Analysis of speech-based speech transmission index methods with implications for nonlinear operations., *J. Acoust. Soc. Am.*, 2004.

[5] Kollmeier, B. et al.: Real-time multiband dynamic compression and noise reduction for binaural hearing aids., *J. of Rehabilitation Research and Development*, 1993.

[6] Kollmeier, B. and Koch, R.: Speech enhancement based on physiological and psychoacoustical models of modulation perception and binaural interaction., *J. Acoust. Soc. Am.*, 1994.

[7] Litovsky, R. Y. et al.: The precedence effect., *J. Acoust. Soc. Am.*, 1999.

[8] Martin, K. D.: Echo suppression in a computational model of the precedence effect., *IEEE ASSP Workshop on Applications of Signal Processes to Audio and Acoustics*, 1997.

[9] Meddis, R. et al.: Implementation details of a computation model of the inner hair-cell/auditory-nerve synapse., *J. Acoust. Soc. Am.*, 1990.

[10] Noisternig, M. et al.: A 3d ambisonic based binaural sound reproduction system., *AES Conf. Paper*, 2003.

[11] Palomaeki, K. J. et al.: A binaural processor for missing data speech recognition in the presence of noise and small-room reverberation., *Speech Communication*, 2004.

[12] Wang D. and Brown, G. J.: Computational Auditory Scene Analysis. *A John Wiley & Sohns, Inc., Publication*, 2006.

[13] Wittkop, T. et al.: Speech processing for hearing aids: noise reduction motivated by models of binaural interaction., *Acta Acustica united with Acustica*, 1997.