Monaural and binaural benefit from early reflections for speech intelligibility

Iris Arweiler, Jörg M. Buchholz, Torsten Dau

Centre for Applied Hearing Research, 2800 Kgs. Lyngby, Denmark, Email: ia@elektro.dtu.dk

Introduction

The auditory system takes advantage of early reflections (ER's) in a room by integrating them with the direct sound (DS) and thereby increasing the effective speech level [1]. The energy, spectral content and direction of the ER's are dependent on wall absorptions and room geometries and therefore different from the DS. By using an ER pattern from a real room and reproducing it in a loudspeaker-based virtual auditory environment, the important characteristics of the ER's can be preserved and the benefit from ER's for speech intelligibility (SI) can be quantified. In the present study SI was measured in such a realistic sound field by varying the level of the ER's and the DS of the speech signal independently. The efficiency of the ER's was then defined as the ratio of the useful ER energy and the total ER energy at the speech reception threshold (SRT). Furthermore monaural SI measures were performed to investigate if ER processing is a monaural or binaural effect.

Methods

The SI measurements took place in an acoustically dampened room with 29 loudspeakers arranged symmetrically around the listener. The room impulse response (RIR) used to create a realistic sound field was taken from a class room modelled with the room acoustic software Odeon [2]. ER's up to the second order were considered, realizing a 55 ms long RIR with 20 reflections. SI was measured monaurally and binaurally with the Danish sentence test Dantale II [3]. The interferer was a diffuse stationary speech shaped noise (SSN) created from the sentence material and presented at a fixed level of 60 dB SPL. For the monaural SI measurements, one ear at a time was closed with an ER2 insert earphone (Etymotic Research). In addition, white noise was presented through the earphone with a level of 75 dB SPL.

Three conditions were tested. In the first condition, only the DS of the speech was presented from 0° azimuth (DS$_{only}$). In the second condition, the DS of the speech was presented from 0° azimuth together with the spatially distributed ER's (DSER$_{spatial}$) provided by the Odeon software. Finally, in the third condition, both the DS of the speech and the ER's were presented from 0° azimuth (DSER$_{frontal}$). In condition one, the level of the speech signal was varied by changing the level of the DS. In conditions two and three, the DS level was kept constant and the level of the spatial or frontal ER's was varied. All SI scores were measured relative to each individual listener's SRT. At the SRT, the contribution of spatial ER's to the overall speech level was 6 dB. The reference point for all conditions was set to the speech level at the SRT minus 6 dB (i.e. no ER's). From this reference point, the SNR was increased stepwise by either adding DS energy or ER energy. For each SNR, the SI was measured with 10 sentences per listener. Nine normal-hearing listeners participated in the experiment.

Results and discussion

Figure 1 shows the mean SI scores with error bars indicating ±1 standard deviation. A logistic function p(SNR) was fit to the data given by:

$$p(SNR) = \frac{1-\alpha}{1+exp(4 \cdot s_{55} \cdot (SRT_{55} - SNR))} + \alpha \quad (1)$$

where SNR is the signal-to-noise ratio, α is the chance level of 10%, SRT_{55} is the SNR at 55% correct SI and s_{55} is the slope at SRT_{55}. There was no significant difference between the left and the right ear, therefore only the results for the left ear are shown. SI was significantly higher for the DS$_{only}$ condition (circles) than for the DSER$_{frontal}$ (diamonds) and the DSER$_{spatial}$ (squares) condition.

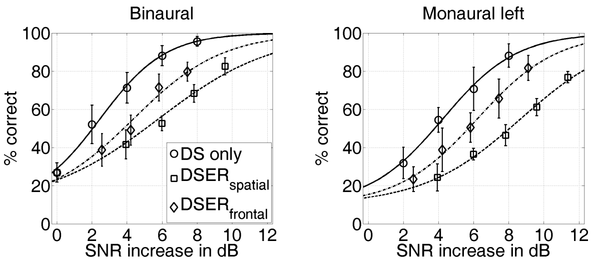

Figure 1: Mean SI scores and fitted psychometric functions for binaural listening (left panel) and listening with the left ear only (right panel).

Intelligibility-weighted SNR

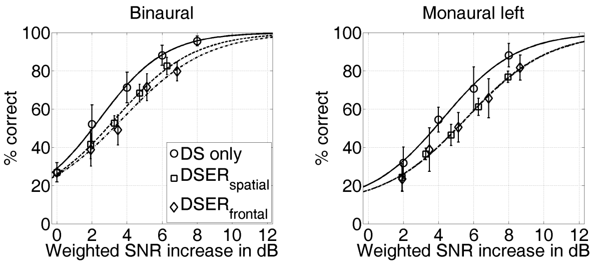

Figure 2: Same as Fig. 1 but for intelligibility-weighted SNR.

Due to the absorptive characteristics of the walls in the simulated class room, the ER's had a different spectrum than the direct sound. In consequence, the ER's might have added energy in frequency regions that are less important for SI. In order to compensate for this spectral effect, the intelligibility-weighted SNR [4] was applied. The speech and noise signals were split into 1/3 octave bands. The SNR was calculated in each band and, before summing, each band was weighted according to its contribution to SI. The band importance function used was taken from table 3 of the SI index (SII) standard [5]. Figure 2 shows the SI scores as a function of the intelligibility-weighted SNR increase. The SI for the DS$_{only}$ condition remained the same, because the spectra of the speech signal and the interferer were identical (the interferer was constructed from the sentences). The SNR for the DSER$_{spatial}$ and DSER$_{frontal}$ conditions decreased compared to the unweighted SNR, indicating that the ER's added energy in frequency regions that are unimportant for SI. The weighted SNR decreased more for spatial ER's than for frontal ER's. The directional filtering of the pinna might have contributed to decreased energy in frequency regions important for SI for spatial ER's compared to frontal ER's. SI for the DS$_{only}$ condition was still significantly better than for the conditions with ER's for weighted SNR increases above 4 dB, which might be due to temporal integration characteristics of the ER's with the DS.

Efficiency factor

In order to quantify the benefit from ER's, an efficiency factor eff_{ER} was introduced, i.e.:

$$eff_{ER} = \frac{\Delta P^*_{ER}}{\Delta P_{ER}} \qquad (2)$$

where ΔP_{ER} is the physical increase of the total speech power by adding the ER's and ΔP^*_{ER} is the corresponding usable power increase. If all the power of the ER's could be used for SI then $eff_{ER} = 1$. If only part of the energy of the ER's could be used for SI then $0 < eff_{ER} < 1$. The *total usable* sound power can then be described as:

$$P_{total} = P_0 + eff_{ER} \cdot \Delta P_{ER} \qquad (3)$$

where P_0 is the power of the DS alone. Eq. 1 was used to estimate eff_{ER}. For SRT55 and s55 the estimates from the DS$_{only}$ condition were used. The SNR corresponds to P_{total} and was varied for $0 < eff_{ER} < 1$ until the minimum root mean square error was found. The estimated eff_{ER} are shown in Fig. 3 for unweighted and weighted SI. For the unweighted SNR, about 25% of the spatial ER energy and up to 52% of the frontal ER energy could be used for SI. Applying the weighted SNR increased the usable energy up to 73%.

Monaural vs. binaural ER processing

Binaural SI was about 2 dB better than monaural SI for all three conditions (see Fig. 1). No binaural speech cues were available in the DSER$_{frontal}$ condition, but there was still a 2 dB improvement. Therefore, the integration of ER's with the DS must have taken place monaurally, and the binaural benefit of 2 dB can be explained by a phase-sensitive summation of the signals at the two ears (i.e., the speech signal is added in phase and the diffuse noise is added in random phase, ideally providing a 3-dB benefit). Since the diffuse noise might have suppressed the full utilization of the binaural system, additional SI measures were performed with a fixed interferer at 90° azimuth and using stationary SSN, multi-talker babble and two "reversed female" talkers. Although the absolute binaural benefit between the three fixed interferer conditions was different, there was no difference in binaural benefit between the DSER$_{spatial}$ and the DSER$_{frontal}$ condition.

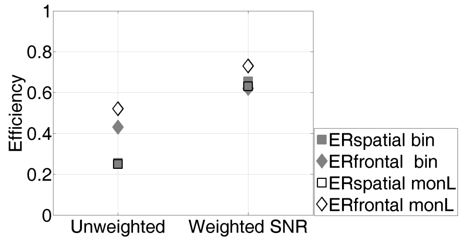

Figure 3: Efficiency factor eff_{ER} for unweighted and weighted SNR at 55% SI.

Conclusions

- Increased ER energy improved SI, but the improvement was less than for increased DS energy. It is important that the ER's used are not just delayed and attenuated copies of the DS, but contain spectral information as in a real room.
- An efficiency factor eff_{ER} was introduced to describe the efficiency of ER's for SI. The efficiency is less for ER's than for DS because of the spectral, spatial and temporal characteristics of the RIR.
- No binaural processing of ER's other than a summation of the signals at the two ears was found. Thus, it is assumed that the integration of the ER's with the DS takes place at an early monaural stage of the auditory system and that the combined signal is then processed binaurally.

References

[1] Bradley, J.S., Sato, H., and Picard, M.: On the importance of early reflections for speech in rooms. J. Acoust. Soc. Am. (2003), 113(6), 3233-3244

[2] Odeon Room Acoustic Software (2008) Version 9.1.

[3] Wagener, K.: Design, optimization and evaluation of a Danish sentence test in noise. Int. J. Aud. (2003), 42, 10-17

[4] Greenberg, J., Peterson, P. and Zurek, P. M.: Intelligibility-weighted measures of speech-to-interference ratio and speech system performance. J. Acoust. Soc. Amer. (1993), 94(5), 3009-3010

[5] ANSI: Methods for the Calculation of the Speech Intelligibility Index. American National Standard S3.5-1997 (1997). Acoustical Society of America