Spatial release from masking for sentence recognition in noise

Juan-Pablo Ramirez[1], Maxime Kolly[2], and Alexander Raake[1]

1 Quality and usability Lab, Deutsche Telekom AG Laboratories, Berlin Institute of Technology, Berlin, Germany

2 Ecole Nationale Supérieure de l'Electronique et de ses Applications, Cergy-Pontoise, France

The scope of this study is to investigate the impact of a single interfering noise on the binaural intelligibility of sentences. Speech Reception Thresholds (SRT) were assessed using semantically unpredictable short sentences presented in the horizontal plan at an angle x while masked by stationary speech shaped noise at an angle y. The test procedure and results are detailed and a Binaural Better Band per ear Speech Intelligibility Index (SII_{3b}) is compared to the resulting Spatial Release from Masking (SRfM), showing satisfactory correlation.

INTRODUCTION

Binaural hearing refers to the situation where the acoustic waves reaching the left and the right ear are instantaneously not identical. This is for example the case when a sound source is located out of the head's median plan. Interaural Level and Time Differences (ILD and ITD) result respectively from head shadowing and phase differences between the two ears. In most cases, the central system better segregates sources when they are spatially distributed, resulting in the so-called Spatial Release from Masking (SRfM).

Speech Intelligibility (SI) is often expressed by the Speech Reception Threshold (SRT). That is considering a speech Target T and a Masker M, the Target-to-Masker Ratio (TMR) yielding an SI of 50%. Bronkhorst (2000) proposed an equation that fits in a simple parametric fashion the release from masking when maskers are located at various angles azimuth. The predictions were based on sets of SRT data proposed by Bronkhorst *et al*. (1992), Plomp *et al*. (1981) and Pessig *et al*. (1997). If the target is not presented from front, Bronkhorst (2000) proposes to adapt his model by considering both relative and absolute masker positions, meanwhile remarking that this assumption only relies on experimental data. The lack of speech intelligibility test data for normal hearing subjects reported and for T elsewhere than in front motivated the present study. The SRT was assessed for a single speech source masked by a single stationary noise, both presented respectively at angles x and y azimuth, yielding to multiple {Tx, My} configurations.

Proceedings of ISAAR 2009: Binaural Processing and Spatial Hearing. 2nd International Symposium on Auditory and Audiological Research. August 2009, Elsinore, Denmark. Edited by J. M. Buchholz, T. Dau, J. Cristensen-Dalsgaard, and T. Poulsen. ISBN: 87-990013-2-2. The Danavox Jubilee Foundation, 2010.

The observed SRfM shows that not only the relative but also the absolute angles impact the listener's ability to segregate sources. Measurements were compared to predictions obtained with the Speech Intelligibility Index (SII) according to ANSI (1997) assessed for the best ear. Experimental and predicted scores show strong correlation.

EXPERIMENTAL PROCEDURE AND RESULTS

Conditions and test setup

SRT were enhanced for all {Tx, My} source distributions possible with x and y values at -120°, -90°, -60°, 0°, 60°, 90°, 120° and 180°. 0° is the front location; negative angles refer to positions at left azimuth, positive angles to the right. Recognition performances were supposed equal at both ears and symmetrical combinations were not presented. Example: {T-60°, M0°} and {T60°, M0°} are supposed to be equivalent, and the later combination only was assessed. Thus 37 position couples were evaluated instead of 64 (see Table 1).

24 young students screened for normal hearing using pure tone audiometry were trained and paid to take part in the test. Each subject provided an SRT measurement at {T0°, M0°} and at 12 different {Tx, My} distributions. Each combination of angles for target and masker was assessed by 8 subjects, apart from {T0°, M0°} which was assessed 24 times.

A corpus of 288 semantically unpredictable sentences of 4 key words each was used. They were uttered in German by a professional male speaker without any particular accent, and recorded in an anechoic chamber. Ramirez *et al*. (2009) further describes the procedure followed to develop the corpus. A speech-shaped stationary noise was created by overlapping 10 times in a randomised fashion the 288 sentences of the corpus arranged end-to-end on a 20 seconds-long wave file. The generated stationary noise M was normalized to a root-mean-square of -32 dBov (relative to digital overload value), and presented at 70 dB SPL (relative to 20 µPa). For each sentence of the corpus, the active speech was extracted via a simple voice activity detection (based on a -50 dBov threshold value of the target speech) and its level was normalized to -32 dBov. Spatial {Tx, My} conditions were generated by convolving target and distracting full-band signals with HRTF. Sources were sampled at 44.1 kHz.

T's level was adjusted by adaptive steps according to the SI of the sentence previously retrieved, targeting a 50 % word recognition score over 12 sentences. The level of the first sentence was set by the user himself. The SRT was assessed as the average level presented across the last 8 sentences. The adaptation-type SRT measurement is based on Plomp (1986), and the employed level step size adaptation is based on Brand *et al*. (2002).

Results

All average SRT values measured and their standard deviation across 8 listeners are given in Table 1.

x \ y	0°	60°	90°	120°	180°
0°	0	9.86	7.26	9.36	2.29
60°	5.21	-0.97	0.49	1.17	9.00
90°	5.27	2.00	-1.01	3.38	8.58
120°	3.85	-0.84	0.18	0.83	7.07
180°	-1.30	7.64	6.52	8.43	-0.81
-120°	5.96	12.22	9.95	12.23	NaN
-90°	6.48	13.08	10.60	12.61	NaN
-60°	5.99	12.15	11.53	13.56	NaN

Table 1: SRfM from the test. The values indicate the SRT measured in dBfp (relative to SRT at frontal position {T0°, M0°} which was -3,2 dB). x refers to the target location, y to the masker location. Positions that read NaN were not measured, as they refer the redundant combinations of x and y. We assume for example that (x,y)=(-90°,180°) is equivalent to (x,y)=(90°,180°).

Subjects reported an average frontal SRT {T0°, M0°} of -3.2 dB. The SRfM values analysed in the following and reported in Fig. 1 refer to this level ratio, and will be given in dBfp (relative to SRT at frontal position {T0°, M0°}). Standard deviation (Std) ranges from 0.48 to 1.55. In general, Std increases with the SRfM. At a given angle θ we observe an average correlation of 0.93 between SRfM {Tθ, My} and SRfM {Tx, Mθ}, for a significant average root mean squared error of 2.2. Thus, release from masking is noticeably modified when target and masker location are inverted. This is particularly pronounced for θ = 0°.

In accordance with the observations presented by Pessig *et al*. (1997), a local minimum of up to 3 dBfp in the SRfM at {Mx, T±90°} appears for all values of x. This is not the case at {M±90°, Ty}. This coincides with the so-called bright spot location. Symmetrically diffracted waves of the masker arrive in phase at the best ear from the opposite side of the head, locally increasing M's energy. Further explanation and modelling of this phenomenon can be found in Avendano *et al*. (1999).

BINAURAL BETTER CRITICAL BAND SII_{3b}

A method to explain and predict the advantage of binaural over monaural listening was proposed by Durlach (1963, 1972), under the name of Equalization-Cancellation (EC) theory, initiating a prolific amount of research in the domain. Among others, two models of SRfM were recently proposed, one using the EC theory coupled with the SII (Beutelman *et al*., 2006), the other relying on simplified binaural cues coupled with the STI (van Wijngaarden *et al*., 2008). Both studies lead to good predictions, but rely on computationally complex algorithms, which were verified on limited

amounts of spatial distributions. They both have the advantage of predicting SRfM in reverberation conditions. In the following, we propose a simple model based on the HRTF energetic properties only.

SRfM varies from -1.30 dBfp to 13.56 dBfp in the study (0 dBfp to 13 dBfp were reported by Bronkhorst (2000). For speech presented at normal loudness (65 dB SPL), this range of levels coincides precisely with the linear dynamic range of the traditional SII described in ANSI (1997).

Experiments performed by Hawley *et al*. (2004) showed that in the case of a single interferer, the better ear accounted for most of the source separation. The traditional SII described by ANSI (1997) was adapted to binaural hearing as follows:

- The long term spectra of both T and M wave files are spatially distributed by filtering through HRTF (same as used in the test). The SII is computed at each ear. In each critical band *i*, the highest of the two ears' SII_i is considered. Eq. 1 in ANSI (1997) is thus transformed as follows:

$$SII_{3b} = \sum_{i=1}^{21} I_i \max(A_{i,left}, A_{i,right}) . \qquad \text{(Eq. 1)}$$

- SII_{3b} {T0°, M0°} is calculated for T and M both in frontal position for TMR = 0 dBfp.
- The SRfM is predicted by the relative variation of speech level yielding the 50% intelligibility score in frontal position, simulating an SRT-type experiment with steps as small as precision is required. In other words T's level is adjusted to have SII_{3b} {Tx, My} = SII_{3b} {T0°, M0°}. This level adjustment is equal to the SRfM in dBfp.

An alternative to this approach consists in measuring the SII_{3b} {Tx, My} at a level of TMR=0 dBfp, and applying a linear fitting on the experimental results, as follows:

$$\text{SRfM} = 41 \cdot \text{SII}_{3b}\{\text{Tx, My}\} - 17 . \qquad \text{(Eq. 2)}$$

With both methods, Pearson correlation is 0.97, and the root mean squared error is 1.19.

Advantages of the SII_{3b}

A first advantage of the approach proposed is that predictions correlate with the observation with similar accuracy as reported in Beutelmann *et al.* (2006) and van Wijngaarden *et al.* (2008), while requiring less computation power. Predictions are base on an exhaustive set of target speech and distracter noise locations, with T being displayed not only in frontal position but all around the head.

Disregarding the alternative proposed in Eq. 2, the approach does not introduce any fitting parameters. It relies only on the spectral features of acoustical sources and HRTFs. Consequences of the bright spot phenomenon are reflected on the intelligibility scores (see Fig. 1b), accounting for a local decrease of the SRfM of 3 dB when M is located in front of the "worst" ear.

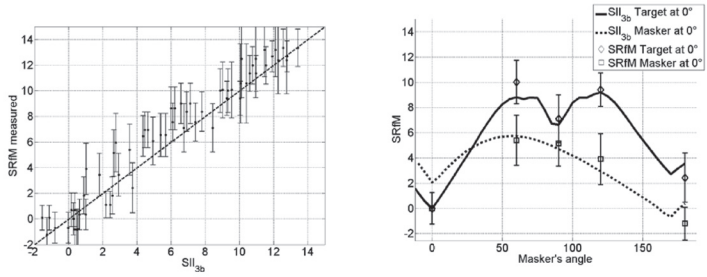

Fig. 1: Experimental SRfM and SII_{3b} based predictions correlation is shown in (a). The model shows robust correlation, but is in general too pessimistic. In (b), solid (resp. dashed) line refers to T (resp. M) in front, and M (and resp. T) at various positions. Diamond and squares are experimental observations with their associated 95% confidence interval, lines are predictions. The bright spot clearly shows in the reported intelligibility.

Limitations of the SII_{3b}

Drawbacks are visible in the accuracy of the prediction for several couples {Tx, My}, although most of them remain in the 95% confidence interval. Predictions are mostly too pessimistic, probably as a consequence of the advantage in source separation resulting from ITDs. The latter will be investigated in future modelling.

Conditions with reverberation or non-linear distortion are not considered in the present model due to the limited range of application of the SII. In order to extend the present model to reverberation, it is proposed to sum the Binaural Masking Level Differences presented in Lavandier *et al.* (2009). Additional intelligibility measurement are however required to validate this hypothesis. Multiple talker scenarios are to be investigated too, as well as the introduction of cues resulting in informational masking.

CONCLUSION

An exhaustive set of measurements of speech reception threshold was performed on a corpus of German semantically unpredictable short sentences masked by stationary speech shaped noise. Both sources were located in the horizontal plan with a combination of an exhaustive set of angles. It showed that inverting target and masker sources did not lead to similar segregation, thus that the absolute location of both sources is to be considered when enhancing spatial release from masking.

As a starting point to more complete binaural modelling, a simple binaural better band per ear SII_{3b} is compared to the observed scores, delivering accurate predictions notably by accounting for the impact of the bright spot on intelligibility.

Combination with models including reverberation and interaural time differences is under study in order to enlarge the field of application of the model.

REFERENCES

ANSI (**1997**). "Methods for calculation of the speech intelligibility index," ANSI Report No. S3.5-1997 (American National Standards Institute, New York).

Avendano, C., Duda, R. O., and Algazi, V. R. (**1999**). "Modeling the contralateral HRTF", 16th International Conference: Spatial Sound Reproduction, Rovaniemi, Finland, 313-318.

Beutelmann, R., and Brand, T. (**2006**). "Prediction of speech intelligibility in spatial noise and reverberation for normal-hearing and hearing-impaired listeners," J. Acoust. Soc. Am. **120**, 332-342.

Brand, T., and Kollmeier, B. (**2002**). "Efficient adaptice procedures for threshold and concurrent slope estimates for psychophysics and speech intelligibility tests," J. Acoust. Soc. Am. **111**, 2801-2810.

Bronkhorst, A. W. (**1992**). "Effect of multiple speechlike maskers on binaural speech recognition in normal and impaired hearing," J. Acoust. Soc. Am. **92**, 3132-3138.

Bronkhorst, A. W. (**2000**). "The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions," Acustica **86**, 117-128.

Hawley, M. L., Litovsky, R. Y., and Culling, J. F. (**2004**). "The benefit of binaural hearing in a cocktail party: Effect of location and type of interferer," J. Acoust. Soc. Am. **115**, 833-843.

Lavandier, M., and Culling, J. (**2008**)."Speech segregation in rooms: Monaural, binaural, and interacting effects of reverberation on target and interferer," J. Acoust. Soc. Am. **123**, 2237-2248.

Pessig, J., and Kollmeier, B. (**1997**). "Directivity of binaural noise reduction in spatial multiple noise-source arrangements for normal and impaired listeners," J. Acoust. Soc. Am. **101**, 1660-1670.

Plomp, R., and Mimpen, A. M. (**1981**). "Effect of the orientation of the speaker's head and the azimuth of a noise source on the speech reception threshold for sentences," Acustica **48**, 325-328.

Plomp, R. (**1986**). "A signal-to-noise ratio method for the speech-reception SRT of the hearing impaired," J Speech Hear Res. **29**, 146-154.

Ramirez, J-P., Raake, A., and Reusch, D. (**2009**). "Intelligibility assessment method for semantically unpredictable sentences in German," Proceedings of the Dega Annual Conference, Rotterdam, Nederland.

van Wijngaarden, S. J., and Drullman, R. (**2008**). "Binaural intelligibility prediction based on the speech transmission index," J. Acout. Soc. Am. **123**, 4514-4523.