ing", in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing Conf. on Acoustics, Speech, and Signal Processing*, 4574–4577.

Kim, C. and Stern, R. M. (**2012**). "Power-normalized cepstral coefficients (PNCC) for robust speech recognition", IEEE Trans. on Audio, Speech, and Language Proc. (accepted for pubication) .

Kingsbury, B. E. D., Morgan, N., and Greenberg, S. (**1998**). "Robust speech recognition using the modulation spectrogram", Speech Communication **25**, 117–132.

Kleinschmidt, M. (**2003**). "Localized spectro-temporal features for automatic speech recognition", in *Proc. Eurospeech*, 2573–2576.

Lyon, R. F. (**1982**). "A computational model of filtering, detection and compression in the cochlea", in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 1282–1285 (Paris).

Mesgarani, N., Slaney, M., and Shamma, S. A. (**2006**). "Discrimination of speech from nonspeech based on multiscale spectro-temporal modulations", IEEE Trans. on Audio, Speech, and Language Proc. **14**, 920–929.

Moore, B. C. J. (**2003**). *An Introduction to the Psychology of Hearing*, fifth edition (Academic Press, London).

Moreno, P. J., Raj, B., and Stern, R. M. (**1996**). "A vector taylor series approach for environment-independent speech recognition", in *Proc. IEEE Int. Conf. Acoust., Speech and Signal Processing*, 733–736.

Pickles, J. O. (**2008**). *An Introduction to the Physiology of Hearing*, 3 edition (Academic Press).

Rabiner, L. R. and Juang, B.-H. (**1993**). *Fundamentals of Speech Recognition* (Prentice-Hall).

Ravuri, S. (**2011**). "On the use of spectro-temporal features in noise-additive speech", Master's thesis, University of California, Berkeley.

Seneff, S. (**1988**). "A joint synchrony/mean-rate model of auditory speech processing", J. Phonetics **15**, 55–76.

Tchorz, J. and Kollmeier, B. (**1999**). "A model of auditory perception as front end for automatic speech recognition", J. Acoustic. Soc. Amer. **106**, 2040—-2060.

Wang, K. and Shamma, S. A. (**1994**). "Self-normalization and noise-robustness in early auditory representations", IEEE Trans. on Speech and Audio Processing **2**, 421–435.

Yost, W. A. (**2006**). *Fundamentals of Hearing: An Introduction*, 5 edition (Emerald Group Publishing).

Young, E. D. and Sachs, M. B. (**1979**). "Representation of steady-state vowels in the emporal aspects of the discharge patterns of populations of auditory-nerve fibers", J. Acoustic. Soc. Amer. **66**, 1381–1403.

Zhang, X., Heinz, M. G., Bruce, I. C., and Carney, L. H. (**2001**). "A phenomenologial model for the response of auditory-nerve fibers: I. nonlinear tuning with compression and suppresion", Journal of the Acoustical Society of America **109**, 648–670.

Modelling the combined effect of binaural hearing and reverberation

Birger Kollmeier[1,2], Jan Rennies[2], Anna Warzybok[1] and Thomas Brand[1]

[1] *Centre for Hearing Research, Medizinische Physik, Universität Oldenburg, D-26111 Oldenburg, Germany*

[2] *Fraunhofer Project group for Hearing, Speech and Audio Technology, Marie-Curie-Str. 2, D-26129 Oldenburg, Germany*

To study the interaction between the intelligibility advantage in rooms due to the presence of early reflections and due to the binaural blocking of interferers from undesired directions, a series of speech reception threshold (SRT) experiments was performed in a simulated room and with a single early reflection of the frontal target speech source as a function of its delay ranging from 0 to 200 ms. From the data and the model considerations given here, one can conclude that binaural unmasking and temporal integration of reflections seem to be comparatively independent from each other, thus providing evidence for a model with a binaural processing stage as a frontend and a reverberation compensation stage (like the MTF model) as the subsequent, independent stage. However, a blocking effect was found for reflections ipsilateral to the noise direction and a release from the deterioration effect at 200 ms delay was found for all non-blocked reflections from azimuths deviating from the midline. These findings are at odds with three versions of a model of binaural speech intelligibility in rooms described here.

INTRODUCTION

Modelling binaural speech reception in normal-hearing and hearing-impaired listeners is a challenging, not yet satisfactorily resolved task especially if complex acoustical environments are involved that are characterized by reverberation and several interfering sound sources. Until now, only three subproblems have been addressed in a satisfactory way:

a) Monaural (i.e., single receiver) speech intelligibility prediction with the combined effect of reverberation and noise has been considered in the Speech Transmission Index (STI-)approach (Steeneken and Houtgast, 1980) and its further developments.

b) Binaural speech intelligibility prediction under nonreverberant conditions (i.e., vom Hövel, 1984, Peissig and Kollmeier, 1997, Beutelmann and Brand, 2006) assuming a simple binaural processing mechanism (i.e., the equalization-cancellation (EC) theory by Durlach, 1972) acting as an optimized two-microphone

Proceedings of ISAAR 2011: Speech perception and auditory disorders. 3rd International Symposium on Auditory and Audiological Research. August 2011, Nyborg, Denmark. Edited by T. Dau, M. L. Jepsen, J. Cristensen-Dalsgaard, and T. Poulsen. ISBN 87-990013-3-0. EAN 9788799001330. The Danavox Jubilee Foundation, 2012.

array which can steer the main lobe and direction of maximum attenuation in the azimuthal plane in a way which optimizes the respective signal-to-noise ratio.

c) The beneficial effect of the direct sound and early reflections in a room (i.e., the first 40 to 100 ms of the room impulse response) on speech intelligibility and the negative or masking effect of the later, spatially diffuse portion of the room impulse response (which is caused by presenting, highly delayed and diffuse portions of the speech that are largely uncorrelated with the direct sound) have been described extensively in the early literature on subjective room acoustics (e.g., Lochner and Burger, 1964, see Kuttruff, 2009 for a review). Arweiler and Buchholz (2011) demonstrated that the spatial distribution of the early reflection component has a (limited) effect on speech reception thresholds.

Even though first attempts to apply a binaural processing model (according to b)) with or without an STI-approach (according to a)) have been quite successful (van Wijngaarden and Drullman, 2008; Beutelmann *et al*., 2010, Lavandier and Culling, 2010), all these approaches did not differentiate between early and late components of the room impulse response (according to c)).

The current contribution therefore investigates the preconditions for a more comprehensive model which is able to predict both the relative aspects of early and late reflections and the role of binaural unmasking for speech intelligibility in rooms in a correct way. The specific research question is: How independent are reverberation integration and binaural unmasking?

NECESSARY EXTENTIONS OF THE BEUTELMANN MODEL

Rennies *et al*. (2011) challenged the binaural speech intelligibility model (BSIM) of Beutelmann *et al*. (2010) (i.e., the combination of an EC-binaural noise reduction and an SII-based speech intelligibility model depicted in Fig. 1c) by measuring speech reception thresholds (SRT) in a virtual room with the conditions given in Fig. 1a): In a virtual reverberant room (T_{60} about 2 seconds) with 8 normal-hearing listeners, SRTs where obtained using the Oldenburg sentence test with the conditions that speech always came from the front (0°) whereas the noise source (steady-state speech-simulating noise) came either from 0°, 22.5° or 90°. Four different distances between speech source and listeners were selected, i.e. 0.5 m, 1.5 m, 3.5 m, and 13.0 m. Their resulting SRTs (Fig. 1b) indicate that with increasing speaker-listener distance the binaural gain decreases: The release from masking when the noise comes from a different direction than the target decreases from approximately 6 dB (for a speaker-listener distance of 0.5 m in condition 1) to approximately 1 dB for condition 4, distance 13 m). This indicates that the room impulse response has an influence on the masker by gradually turning the (directed) lateral masker into a more or less omnidirectional, diffuse masker.

In addition, with increasing speaker-listener distance the SRT increases. This reflects the influence of the room on the speech signal in a way which is well described by the reduced modulation transfer function (i.e. the filling up the “valleys” in the original speech signal by reverberation “tails” and thus decreasing the “effective” speech to noise ratio). When employing the BSIM model of Beutelmann *et al*. (2010) (curves without error bars in the upper panel of Fig. 1b), it becomes clear that this model is able to predict the decrease of binaural unmasking with increasing speaker-listener distance (i.e., the effective decorrelation of the masker). However, it is unable to predict the second effect, i.e. the increase of SRT with increasing speaker-listener distance even if speaker and interferer come from the same direction (condition S_0N_0).

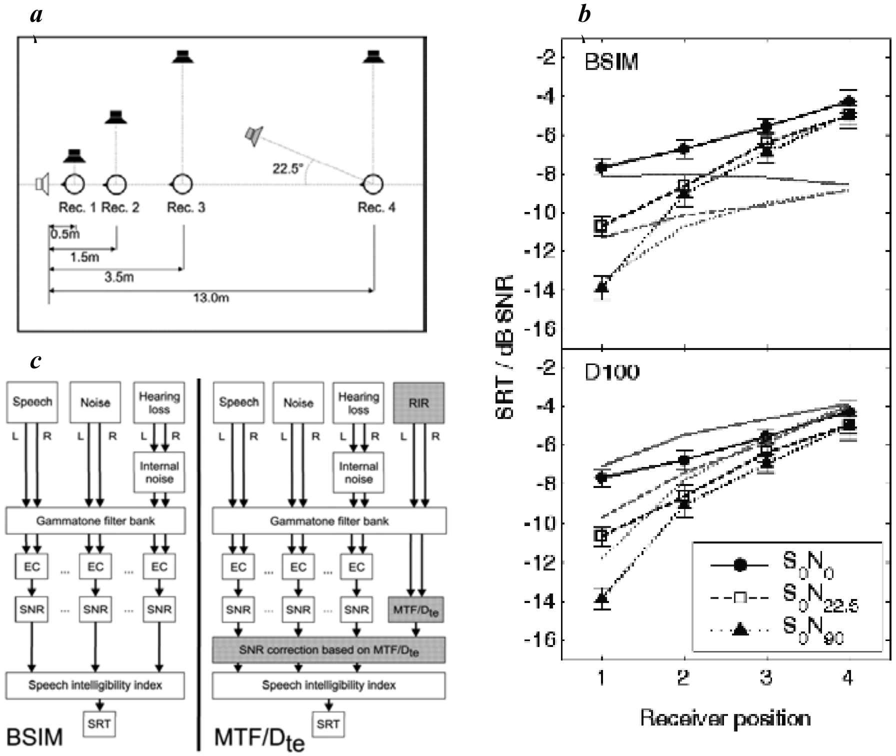

Fig. 1: Simulated spatial configuration (Fig. 1a, left top), measured (including interindividual standard deviation) and modelled speech reception thresholds (Fig. 1b, right) and scheme of the models employed by Rennies *et al*. (2011) (Fig. 1c, left bottom). D100 denotes the model D_{te} with a separation time between early and late portion of the room impulse response (RIR) of 100 ms.

In order to predict the data, Rennies *et al*. (2011) suggested two modifications (Fig. 1c, right panel) of the original BSIM model which utilize properties of the room impulse response (RIR) to provide a better fit to the data:

a) Model MTF: Combination of the binaural frontend from the original BSIM model with a STI-based backend which bases its estimation of the “effective” signal to-noise ratio in each frequency channel not on an instantaneous SNR estimate, but on the speech transmission index in a way proposed by Steeneken and Houtgast (1980). This modification provides a better prediction of the SRT-dependence on speaker-listener distance but tends to overestimate the effect of reverberation, i.e. the SRTs at the conditions 4 were estimated by approximately 3 dB too high. This reason for this overestimation may be due to the fact that the current model explicitly incorporates binaural unmasking whereas the original MTF-based STI did not include any binaural component but estimated the average overall speech intelligibility (including the general binaural effect)is unclear (see Rennies *et al*., 2011).

b) D_{te} ("Definition with transition time te") model D100 which utilizes the room acoustical parameter clarity or "definition" (D_{te}, here: D100) as the ratio of the first 100 ms of the RIR power over the total power of the room impulse response, assuming that the first 100 ms of the room impulse response is the "useful" portion of the RIR that includes the early reflection. This model has been motivated by the early work in room acoustics and assumes that the "effective" speech-to-noise ratio is corrected by increasing the speech by a factor of D_{te} and the noise by a factor of (1- D_{te}). As can be seen from figure 1b (lower panel), this model provides a much better description of the empirical data than the original model version.

In conclusion, the combination of a binaural speech intelligibility model with model components taken from room acoustics appears to be applicable for the conditions shown here. Adding an explicit processing of early reflections improves the potential of the model.

From the success of this combined approach one can postulate that the two mechanisms (i.e. the binaural-processing or noise-blocking mechanism and the (monaural) reverberation processing mechanism integrating early reflections and describing the deterioration effect of late reflections) work independently of each other. This independence hypothesis between binaural processing and reverberation processing was tested in the experiments described below.

INTEGRATION OF A FRONTAL REFLECTION

In order to challenge the models integrating binaural noise reduction and the processing of early reflections and late reverberation, Warzybok *et al*. (2011) performed a series of SRT experiments with 12 normal-hearing listeners with a single early reflection of the frontal target speech source which either originated from the same direction as the target (described in the following) or which originated from different spatial directions (see next section). The interfering noise was either a diotic noise (denoted as N_0), a localized lateral noise at 135 degrees (denoted as N_{135}), or a diffuse noise without a special direction of incidence (denoted N_D, see Fig. 3). In the current experiment, the frontal reflection had the same amplitude as the direct sound and was varied in delay with respect to the direct sound from 0 to 200 ms (0, 10, 25, 50, 75, 100, and 200 ms). All tests were performed via headphones with a standard set of binaural HRTF functions (CATT Acoustics v8.0a) using the Oldenburg sentence test (Wagener *et al*., 1999).

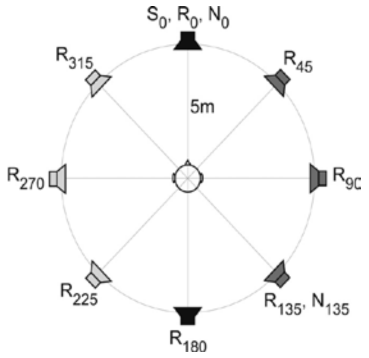

Fig. 2: Spatial configurations for the experiments by Warzybok *et al*. (2011): The direct sound of the speech material was always presented frontally (S_0); noise was also presented frontally (N_0), laterally (N_{135}), or diffusely (N_D, not shown). Black speakers indicate diotic reflections (R_0 and R_{180}), dark gray speakers indicate a reflection from the same side as the lateral noise source (R_{45}, R_{90}, and R_{135}), and light gray speakers indicate a reflection from the opposite direction (R_{225}, R_{270}, and R_{315}). The azimuth of the single reflection varied in the experiments in steps of 45°.

The empirical data from Warzybok *et al*. (2011) are given as solid lines (with error bars indicating interindividual standard deviations) in Fig. 3 a)–c), respectively, whereas the model prediction of the three models outlined above are given as dashed lines. As the delay between direct sound and the single, first reflection increases up to approximately 25 ms, the SRT stays comparatively constant in all three noise conditions indicating a complete integration of the first reflection with the direct sound. With further increasing delay, the the 3-dB integration effect becomes less efficient and vanishes at a delay of approx. 100 ms (i.e., SRT is 3 dB worse than the reference threshold with complete integration). A deterioration effect of the late reflection becomes apparent at a delay of 200 ms since the SRT increases by approximately 5 to 6 dB with respect to the reference condition (0-ms delay). while the detrimental effect of this late reflection is less than 3 dB for delays up to 200 ms. In addition, the binaural intelligibility level difference (BILD), i.e., the difference between condition N_0 and conditions N_{135} and N_D, respectively, is approximately constant for all delays. This provides clear evidence that the integration process of early reflections in the temporal domain operates independently of the binaural, spatial processing.

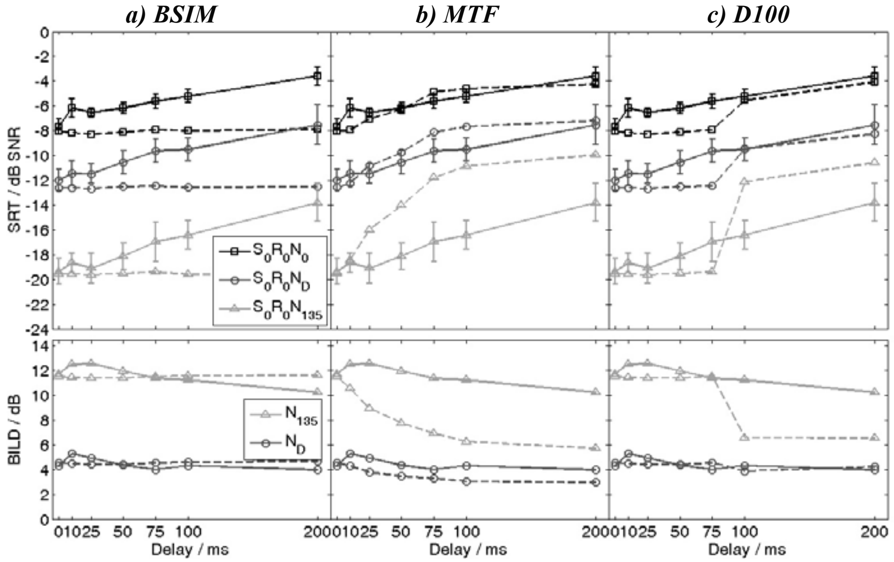

Fig. 3: Top: Mean SRTs (solid lines) and interindividual standard deviations as a function of delay of a frontal reflection of the speech signal in diotic noise (circles), laterally located noise (squares), and diffuse noise (triangles). Direct sound was presented frontally. Each panel gives the same experimental data (solid lines, from Warzybok *et al*., 2011) in comparison with one of three prediction models (dashed lines). Bottom: mean binaural intelligibility level differences for laterally located (squares) and diffuse noise (triangles) as a function of reflection delay.

The model data (given as dashed lines in Fig. 3 for the three models BSIM, MTF and D100) can only partially reproduce the empirical data: The original BSIM (Beutelmann *et al*., 2010) provides a very good prediction of the binaural unmasking effect, thus predicting the BILD quite well, but does not account for the integration of the early reflection. Instead, no dependence of the time difference between direct sound and early reflection is predicted. This is not the case for the MTF-model (middle panel in Fig. 3) which can predict the diotic condition (upper curve) quite well but considerably overestimates the negative effect of the late reflection on the binaural unmasking for the N_{135} condition. Obviously, this model can describe the early reflection integration for the monaural case in an appropriate way, but does not take into consideration that the early reflection seems to be treated by the binaural system as a part of the target signal and not as a part of the masker. Similarly, the third model variant D100 (Fig 3c) is neither able to predict the reflection integration in a correct way nor to account for the binaural interaction in the direct sound (plus reflection) versus the interfering sound: The assumed steep separation at the boundary between direct sound and reverberant tail at 100 ms leads to a very steep transition in the predictions from a model behaviour similar to the original BSIM-model (for small delays) to an MTF-type-model (for delays grater than 100 ms).

As a conclusion, none of the three models employed here can describe the empirically found independence between binaural processing and integration of early reflections in the situation with one frontal reflection with varying delay. The next experiment therefore challenges this “independence hypothesis” in order to find out if a model should be built with a complete independence of binaural processing stage and early reflection integration or if there should be appropriate interactions.

CHALLENGING THE INDEPENDENCE HYPOTHESIS

Integration of a spatially separated reflection in lateral noise (at 135°)

While the direction of the early reflection was held constant from the same direction as the target in the previous section, the experiment described in this section varied the azimuth of the refection in steps of 45° using a fixed lateralized noise source at 135°. A subset of delays (10, 50, and 200 ms) was employed between direct sound and early reflection. Ten out of the twelve normal-hearing listeners from the previous experiments participated in the headphones experiment measurements. The resulting SRTs are given in Fig. 4 (from Warzybok *et al*., 2011).

For the diotic reflection conditions (i.e., $S_0R_0N_{135}$, and $S_0R_{180}N_{135}$, respectively), the same dependence on the delay between early reflection and direct sound is observed as before with an advantage of approximately 1 dB for the reflection coming from the rear instead of coming from the front. This indicates a small, non-significant front-back advantage due to some extra spectral information carried by the reflection from the rear. For the cases of contralateral reflection (i.e. $S_0R_{225/270/315}$ N_{135}, light gray lines), a parallel shift to the diotic case is observed, indicating a similar integration of the early single reflection as in the diotic case, but a slight binaural advantage due to the fact that the early reflection comes from a different side than the interfering noise and hence adds some additional binaural cues that the system can exploit. Interestingly, this binaural unmasking effect also holds for delays up to 200 ms where obviously the deterioration effect due to the late reflection is overruled by the small binaural advantage.

In the ipsilateral reflection case (i.e., conditions $S_0R_{45/90/135}N_{135}$, dark gray curves), a comparatively flat function (i.e., no dependence on the delay between early reflection and direct sound) is observed, indicating the lack of an integration effect but also no deterioration by late reflections. This holds especially for the $S_0R_{135}N_{135}$ condition and to a somewhat lesser degree to the $S_0R_{45/90}N_{135}$ conditions where a slight binaural advantage (in the order of 1-2 dB) is visible. Obviously, the ipsilateral reflection is (at least partially) masked by the noise source before any temporal integration or masking can take place.

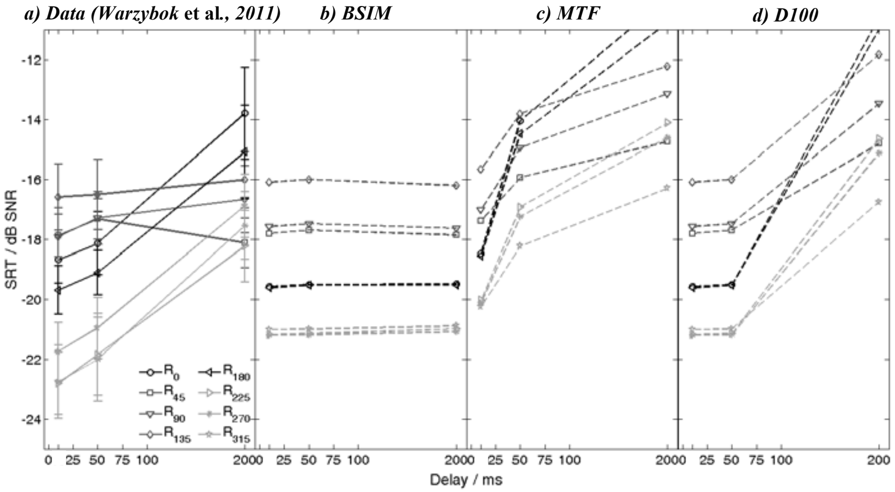

Fig. 4: SRTs measured with frontal speech and a lateral noise source for different reflection azimuths (panel a), from Warzybok *et al*., 2011) and model predictions (panel c)-d)). Black symbols and curves indicate diotic speech signals (reflection from the front or behind), dark gray symbols and curves indicate a reflection azimuth at the same side as the noise source (ipsilateral condition), and light gray symbols and curves indicate a reflection from the opposite direction (contralateral condition).

As an intermediate conclusion, binaural unmasking is not independent from the integration of early reflections and deterioration effect caused by late reflection: In the condition described above, the early reflection is largely masked if the noise comes from the same direction (or the same hemisphere) as the interferer. This condition resembles the observation of Peissig and Kollmeier (1997) who explained their SRTs in the presence of two maskers from different directions by the subjects' inability to cancel out two noise sources from different directions at the same time. Only if both noises come from approximately the same direction (or the same hemisphere), a SRT advantage can be observed.

Figure 4 b)-d) show the prediction performance of the models (dashed lines) outlined above in the same way as the empirical data from Fig. 4 a). Obviously, none of the models can account for the empirical behaviour in the correct way. The BSIM model again does not predict any reflection integration, but predicts the binaural effect approximately correctly, whereas the MTF model does neither predict the reflection integration in a correct way nor the binaural unmasking effect. The same holds for the D100 model because the assumption that a reflection is masked by the noise is not part of the specifications of the model.However, the data for delays below 50 ms better are predicted by the D100 model than by the MTF model.

To test the hypothesis that the interaction between the early reflection and the noise source is influenced by the directional properties of the interferer, the next experiment employs a diffuse noise as a masker.

Challenging the directivity hypothesis of the noise masker: Spatially separated, signal reflection with diffuse noise interferer

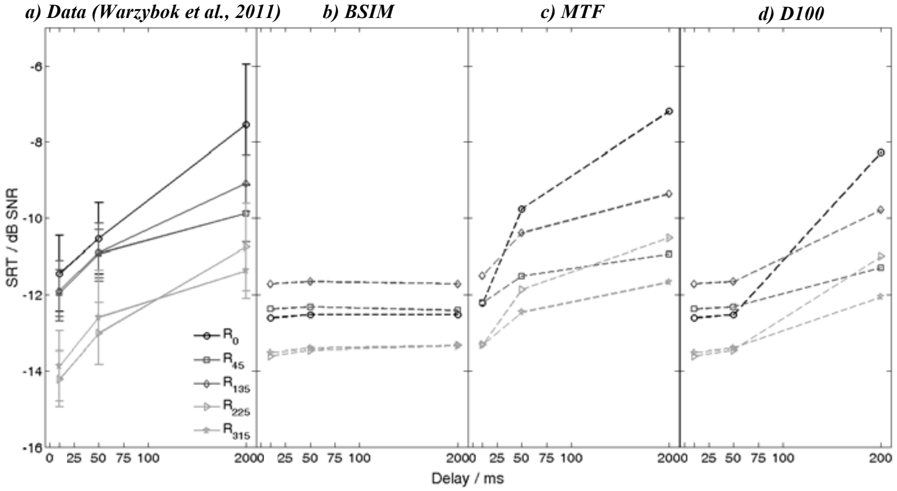

Fig. 5: Same data representation as in Figure 4, but for diffuse noise instead of lateral noise.

The final experiment in this series employs a spatially separated signal reflection varying in delay (10, 50 and 200 ms) and in azimuth of the reflection (0, 45, 135, 225 and 315 degree). As a masker, diffuse noise was employed. The data (together with the corresponding predictions of the three models) are displayed in Fig. 5. In the diotic case (upper curve, black) the same transition between early reflection integration and deterioration due to late reflection is observed as before. In the spatially separated conditions, a slight asymmetry between the left and right ear is observed which is due to the calibration procedure that employs the right ear as reference. Irrespective of these differences in the order of 2 dB, an integration takes place like in the diotic case, but the difference between the smallest delay and the 200-ms condition is less than 3 dB, indicating that a much smaller deterioration effect for the 200-ms reflection is observed if the reflection arrives not from the same direction as the target. This effect is of considerable interest, because an independence hypothesis would predict the same late reflection deterioration for the lateral reflections as for frontal reflections.

As before, the model predictions do not coincide very well with the actual data. While the BSIM model again does not reflect the reflection integration in an appropriate way but gives the correct value for the binaural effect, the MTF model seems to predict both the reflection integration and the binaural effect in an approximately appropriate way. Conversely, the D100 model does not predict the reflection integration appropriately and also does not predict the binaural effect in an appropriate way. From all these models, the MTF model seems to have the overall best performance.

GENERAL CONCLUSIONS

From the data and the first model considerations given here, one can conclude that binaural unmasking and temporal integration of reflections seem to be comparatively independent from each other, thus providing evidence for model with a binaural processing stage as a frontend and a reverberation compensation stage (like the MTF model) as the subsequent, independent stage. However, the following restrictions have to be applied to these models:

- Such an approximate independence between binaural processing and reverberation processing seems to be valid only for the case of the single reflection from the front as discussed in experiment 1 here.
- This independence is definitely not valid for reflections originating from the hemisphere of a localized noise source. In these cases, obviously the reflection is cancelled by the binaural system (in the same way as the noise source) before it can be integrated into a single target object.
- The deterioration effect for long delay times can be cancelled if the reflection comes from a different location than the target. Obviously, the binaurally displayed and temporally resolved reflection can be characterized as a separate object which is neither integrated with the target sound (i.e., no enhancement effect as for the early reflection is observable) nor used as an interferer for the target signal (as would be the case if the reverberation would come from the same direction as the target).

Several consequences can be drawn for models that have to be developed in order to predict the effects described above:

- In general, the models considered here yield a good approximation for more complex reverberation patterns as investigated here (such as e.g. given by Beutelmann *et al*., 2010 and by Rennies *et al*., 2011) and seem to be at their limits in the challenging situations with only one single reflection as considered here.

- The interaction between reverberation and the binaural unmasking effect is not yet described well by these models, since they mostly assume an independence of both components which obviously is not the case in the conditions employed here.

Modification of the available models (like the BSIM and the MTF and models derived from that) should therefore:

- Incorporate binaural unmasking and masking of early reflections before these early reflections are being fused together with a direct sound in a later processing stage.
- Incorporate the reduced deterioration effect by a spatially displaced late reflection. It can be assumed that the later part of the impulse response (i.e., the reverberation "tail") might be perceived as a separate object if it originates from a different direction than the target sound source.

Taken together, the models employed here give a good approximation for complex situations but operate beyond their limits with the comparatively simple, but not very naturalistic situation employed here. Nevertheless, these conditions provide a "critical condition" to uncover the limitations of current models and will therefore help us to develop better models in the future.

ACKNOWLEDGEMENT

Supported by MWK Niedersachsen (stipend to A. W. and Fraunhofer project group support) as well as by SFB/TRR 31 (grant to B. K. and T. B.). The authors wish to thank all subjects for the participations and all members of the medical physics department for their support and help.

REFERENCES

Arweiler, I., and Buchholz, J. M. (**2011**). "The influence of spectral characteristics of early reflections on speech intelligibility" J. Acoust. Soc. Am. **130**, 996-1005.

Beutelmann, R., Brand, T., and Kollmeier, B. (**2010**). "Revision, extension, and evaluation of a binaural speech intelligibility model" J. Acoust. Soc. Am. **127**, 2479-2497.

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**, 331-342.

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

Durlach, N. I. (**1972**). "Binaural signal detection: Equalization and cancellation theory" Foundations of Modern Auditory Theory, edited by J. Tobias (Academic, New York), Vol. II, 371-462.

Kuttruff, H. (**2009**). "Room acoustics" 5th edition, (Taylor & Francis, New York).

Lavandier, M., and Culling, J. F. (**2010**). “Prediction of binaural speech intelligibility against noise in rooms" J. Acoust. Soc. Am. **127**, 387-399.

Lavandier, M., Culling, J. F., and Jelfs, S. (**2010**). “Prediction of reverberant speech intelligibility against multiple noise interferers in rooms: Binaural useful-to-detrimental ratios (A)” J. Acoust. Soc. Am. **128**, 2361.

Lochner, J. and Burger, J. (**1964**). “The influence of reflections on auditorium acoustics” J. Sound Vib. **1**, 426-454.

Peissig, 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.

Rennies, J., Brand, T., and Kollmeier, B. (**2011**). “Prediction of the inuence of reverberation on binaural speech intelligibility in noise and in quiet" J. Acoust. Soc. Am. (in press).

Steeneken, H. J. M. and Houtgast, T. (**1980**). “A physical method for measuring speech-transmission quality" J. Acoust. Soc. Am. **67**, 318-326.

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

vom Hövel, H. (**1984**). “Zur Bedeutung der Übertragungseigenschaften des Außenohres sowie binauralen Hörsystems bei gestörter Sprachübertragung" (On the importance of transmission properties of the outer ear and the binaural auditory system for disturbed speech transmission), doctoral dissertation, RWTH Aachen.

Wagener, K., Brand, T., and Kollmeier, B. (**1999**). “Entwicklung und Evaluation eines Satztests für die deutsche Sprache II: Optimierung des Oldenburger Satztests (Development and evaluation of a German sentence test II: Optimization of the Oldenburg sentence test)" Z. Audiol. **38**, 44-56.

Warzybok, A., Rennies, J., Brand, T., Doclo, S. and Kollmeier, B. (**2011**). “Effects of spatial and temporal integration of early reflections on speech intelligibility” J. Acoust. Soc. Am. (submitted).

Predicting speech intelligibility in adverse conditions: evaluation of the speech-based envelope power spectrum model

SØREN JØRGENSEN AND TORSTEN DAU

Centre for Applied Hearing Research, Technical University of Denmark, DK-2800 Lyngby, Denmark

The speech-based envelope power spectrum model (sEPSM) [Jørgensen and Dau (2011). J. Acoust. Soc. Am., **130** (3), 1475–1487] estimates the envelope signal-to-noise ratio ($\mathrm{SNR}_{\mathrm{env}}$) of distorted speech and accurately describes the speech recognition thresholds (SRT) for normal-hearing listeners in conditions with additive noise, reverberation, and nonlinear processing by spectral subtraction. The latter represents a condition where the standardized speech intelligibility index and speech transmission index fail. However, the sEPSM is limited to stationary interferers due to the fact that predictions are based on the long-term $\mathrm{SNR}_{\mathrm{env}}$. As an attempt to extent the model to deal with fluctuating interferers, a short-time version of the sEPSM is presented. The $\mathrm{SNR}_{\mathrm{env}}$ of a speech sample is estimated from a combination of $\mathrm{SNR}_{\mathrm{env}}$-values calculated in short time frames. The model is evaluated in adverse conditions by comparing predictions to measured data from [Kjems *et al.* (2009). J. Acoust. Soc. Am. **126** (3), 1415-1426] where speech is mixed with four different interferers, including speech-shaped noise, bottle noise, car noise, and cafe noise. The model accounts well for the differences in intelligibility observed for the different interferers. None of the standardized models successfully describe these data.

INTRODUCTION

Models of speech intelligibility can be very useful as tools for investigating which features of the physical speech signal are crucial for understanding the speech in a noisy background. Moreover, an accurate prediction metric is of great relevance in practical applications such as hearing-aid and telecommunication development. Current intelligibility metrics include the articulation index (AI) and its successor the speech intelligibility index (SII). SII-based metrics estimate the effective amount of audible speech information in a number of frequency bands, from the long-term frequency spectra of speech and noise. The audible information is weighted by an empirically determined importance function, describing the relative importance of the individual frequency bands to intelligibility. This approach can predict the intelligibility of speech subjected to low-pass and high-pass filtering and the effects of different stationary noise backgrounds (Kryter, 1962). However, the SII-metric is based on frequency information only, and cannot be successfully applied to conditions with reverberation. As an alternative, the speech transmission index (STI)

Proceedings of ISAAR 2011: Speech perception and auditory disorders. 3rd International Symposium on Auditory and Audiological Research. August 2011, Nyborg, Denmark. Edited by T. Dau, M. L. Jepsen, J. Cristensen-Dalsgaard, and T. Poulsen. ISBN 87-990013-3-0. EAN 9788799001330. The Danavox Jubilee Foundation, 2012.