Speech reception in noise: How much do we understand?

Birger Kollmeier, Bernd Meyer, Tim Jürgens, Rainer Beutelmann, Ralf M. Meyer, Thomas Brand

Medical Physics, Universität Oldenburg, D-26111 Oldenburg, Germany

In order to better understand the effect of hearing impairment on speech perception in everyday listening situations as well as the still limited benefit of modern hearing instruments in this situations, a thorough understanding of the underlying mechanisms and factors influencing speech reception in noise is highly desirable. This contribution therefore reviews a series of studies by our group to model speech reception in normal and hearing-impaired listeners in a multidisciplinary approach using "classical" speech intelligibility models, functional perception models, automatic speech recognition (ASR) technology, as well as inputs from psycholinguistics. Classical speech-information-based models like the Articulation Index or speech intelligibility index (SII) describe the acoustical layer and yield accurate predictions only for average intelligibility scores and for a limited set of acoustical situations. With appropriate extentions they can model more audibility-driven and even time-dependent acoustical situations, such as, e.g. the effect of hearing impairment in fluctuating noise. However, to describe the sensory layer and suprathreshold processing deficits in humans, the combination of a psychoacoustically motivated preprocessing model with a pattern recognition algorithm adopted from ASR technology appears advantageous. It allows a detailed analysis of phoneme confusions and the "man-machine-gap" of approx. 12 dB in SNR, i.e., the superiority of human world-knowledge-driven (top-down) speech pattern recognition in comparison to the training-data-driven (bottom-up) machine learning approaches. Finally, the cognitive abilities of human listeners when understanding speech can be assessed by a "fair" comparison between Human Speech recognition and ASR that employs only a limited set of training data. In summary, both bottom-up and top-down strategies have to be assumed when trying to understand speech reception in noise. Computer models that assume a near-to-perfect "world knowledge", i.e., anticipation of the speech unit to be recognized, can surprisingly well predict the performance of human listeners in noise and may prove to be a useful tool in hearing aid development.

INTRODUCTION

The perception of speech in normal and hearing-impaired listeners is mostly performed under non-ideal, i.e. "difficult" acoustical situations that can only be approximated in the laboratory up to a certain extent. Within these limitations, substantial progress has been made within the last decades to understand speech reception and the specific influence of the various parameters involved. Typically, the speech reception threshold in noise is assessed, i. e., the speech-to-noise ratio required to achieve a cer-

Auditory signal processing in hearing-impaired listeners. 1st International Symposium on Auditory and Audiological Research (ISAAR 2007). T. Dau, J. M. Buchholz, J. M. Harte, T. U. Christiansen (Eds.).
ISBN: 87-990013-1-4. Print: Centertryk A/S.

tain percent correct (in most cases: 50%) of the speech material employed. Auditory models are then used to predict these results and to test the degree to which we quantitatively "understand" human speech recognition in terms of the acoustical information contained in the speech, the sensory processing deficits of the individual listener, and his or her cognitive abilities. The current contribution gives an overview of different modelling approaches for speech reception in noise by reviewing some of the contributions from the Oldenburg labs to these issues within the last years. A rough classification of these models can be given according to their intended level within the communication chain. Hence, the current contribution is organized to address these different layers of speech perception successively.

THE ACOUSTICAL LAYER. ARTICULATION INDEX PLUS EXTENTIONS

The "classical" approach to model speech recognition under noise uses a spectral weighting of the long-term signal-to–noise-ratio and assumes that the total received information is the sum of the information transmitted in different frequency channels where the amount of information in each frequency channel is given by the respective signal-to-noise ratio (Fletcher and Galt, 1950). While a vast literature exists on the articulation index, its further developments (Speech Transmission Index (STI, see Houtgast and Steeneken, 1985) and Speech Intelligibility Index (SII, ANSI, 1997)) and its use for predicting speech intelligibility in hearing-impaired listeners, the following parameters are critical in the attempts to accurately predict the individuals speech reception ability under a certain acoustical condition with or without a certain hearing impairment:

- Shape and weighting of spectral bands: Usually, auditory critical bands (ERB-Scale) with a weighting given by the SII standard (ANSI, 1997) are employed.
- Additivity of external and threshold-simulating noise: To represent the individual hearing loss, a threshold-simulating noise has to be assumed that is spectrally added to the external noise that exists in the respective acoustical situation. In certain situations it seems justified to assume an amplitude additivity of external and internal noise ("coherent addition") rather than the standard additivity of power (see below).
- Short-term segmentation: Even though the AI and most of its historical derivatives were only designed for long-term predictions based on the SNR averaged across large periods of time, a short-time-SNR-based "instantaneous AI" has the advantage of being able to follow time-varying background noise conditions and characteristics of the target signal (c.f. Kollmeier, 1990, Rankovic, 1997, Payton and Brainda, 1999, Wagener and Brand, 2005, Wagener *et al.*, 2006). However, the duration of the time windows for evaluating the SNR, the way of averaging across time and the information combination across instances in time and frequency are critical for the successful prediction of speech recognition. In the current study, either the SNR from the complete utterance

was used for time-independent predictions or window lengths of 30 ms were assumed if indicated.

Methods

A total number of 113 both normal-hearing and hearing-impaired listeners with various degrees of sensorineural hearing loss participated in data collection (Brand and Kollmeier, 2002a). They underwent a clinical tone audiogram, a categorical loudness scaling procedure, and – among other tests – the Oldenburg sentence intelligibility test both in quiet and in noise using an adaptive SRT determination procedure (Brand and Kollmeier, 2002b). The maskers were either an unmodulated ICRA 1 noise (Dreschler *et al.*, 2001) or a modulated ICRA 5-250 noise where the maximum pause duration was limited to 250 msec (Wagener *et al.*, 2006). The noise level was always set to be close to the "medium" categorical unit from loudness scaling.

Results & discussion

Figure 1 shows the obtained speech reception threshold (SRT) in quiet for the Oldenburg sentence test (Wagener *et al.*, 1999) as a function of the predicted SRT. Obviously, the empirical results coincide quite well with the predictions that are derived only from the audiogram (r = 0,95, 69% of all points within the confidence region given by the accuracy of the test).

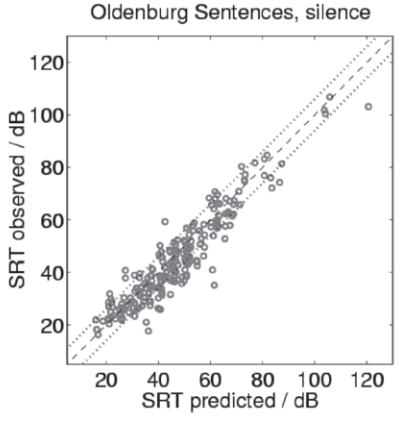

Fig. 1: Speech reception thresholds (SRT) obtained for 113 normal and hearing-impaired subjects with the Oldenburg Sentence test in quiet (ordinate). The predictions obtained with the SII are given on the abscissa (r=0,95).

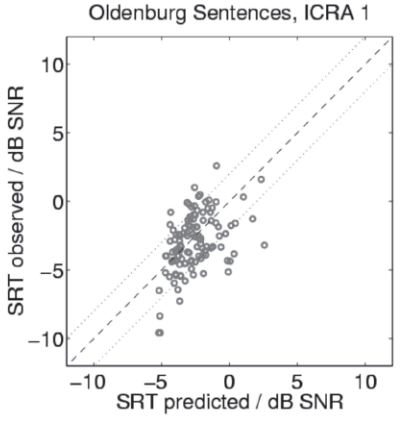

Fig. 2: As Fig. 1, but with continuous speech-shaped background noise (ICRA1) as background. The predictions obtained with the SII are given on the abscissa (r=0,95).

Figure 2 gives the results for the same subjects and the same speech test but measured for stationary ICRA 1 noise. Here, the SRT prediction is no longer dominated by the large variations in the audiogram as in figure 1, but rather comparatively small variability across subjects occur. This variability is not very well predicted by the SII which

is based on the audiogram and the external masking noise. Note that the predictions employed here base on the assumption of a coherent addition between external masking noise and individual threshold-simulating noise. If only the maximum from any of both quantities or an incoherent addition would be used, the prediction would be even worse. The fact that the best predictions are achieved with coherent addition may be justified by the observation, that even a subtle masking noise close to the absolute threshold already has a detrimental effect to signal detection.

Taken together, the time-independent SII reaches a reasonable well prediction accuracy for the SRT in quiet which, however, might not yield much additional information than the audiogram. The situation is different with the supra threshold tests in noise where the limits of the current SII models becomes clear. It is highly probable that the recruitment phenomenon and other suprathreshold processing deficits will be responsible for the observed deviations between empirical SRT data and audiogram- and external noise based SRT predictions. This finding calls for more refined modelling approaches (see below).

As already pointed out above, a great challenge of the prediction models is to correctly predict speech reception thresholds in fluctuating noise and the listener´s ability to "listen into the dips". Fluctuating noise provides a larger difference in SRTs between normal and hearing-impaired listeners. However, the appropriate prediction of the empirical results has not been very satisfactory in the past. We therefore compared four different model versions derived from the SII and implemented them as short-term-predictions based on short segments that vary with centre frequency (Meyer and Brand, 2007):

I.) SII (ANSI, 1997): The starting point is the standard SII which is based on the long-term spectra of speech and noise. The audibility in 21 frequency bands is calculated and the weighted sum (band importance function depending on test material) over all bands is calculated. Consequently, the original time-independent version of the SII is insensitive to temporal fluctuations of the input signals, as the standard is based on the long-term spectra only.

II.) Frequency independent fluctuations of the noise (Brand and Kollmeier, 2002a): In a first step towards a short-term SII, a version proposed by Brand is used which assumes fluctuations of the overall level of the noise whereas the frequency spectrum of the noise is regarded as being constant. For every level occurring in the noise level-histogram an SII value is calculated. Finally the weighted (with the rate in the level-histogram) mean over all SII values is calculated.

III.) Frequency dependent fluctuations of the noise (Rhebergen *et al.*, 2005): In the second step, also the frequency dependency of the fluctuations of the noise are considered. This is done by using the model proposed by Rhebergen *et al.* This model proposes a pre-processing of the input signals where the signals are first filtered into 21 frequency bands. In every frequency band the envelope is estimated via the Hilbert-transform. In frequency dependent time windows the instantaneous intensity is estimated. At last the mean over all SII values is calculated. A noise with the long-term spectrum of speech

is used as representation of the speech signal, as it was done by Rhebergen *et al.* Since this speech simulating noise shows no fluctuations, this approach does not take fluctuations of the speech spectrum into account.

IV.) Frequency dependent fluctuations of speech and noise: In the last step, also the fluctuations of the speech are considered (Meyer and Brand, 2007). This is achieved by taking real speech signals (sentences from the sentence test) as input. For every speech signal the SRT is calculated with the model according to Rhebergen *et al.* (2005) and then the mean over all SRTs is calculated. This requires much more computation time than the other versions of the model. The only difference to model version III) is that speech signals are used as input and that the averaging takes part across much more speech samples.

In each version of the model, every resulting SII value is transformed into an intelligibility value. The speech level is then adjusted to achieve an SII of 0.133. This SII value corresponds to the Speech Reception Threshold (SRT). The subject's hearing-loss is included in the SII as described in the standard.

Figure 3 shows scatter plots of the results for all model versions used. On the abscissa the predicted SRT values are shown. On the ordinate the observed SRT values are shown. The solid diagonal lines represent perfect predictions of the measured data. The dashed lines around the diagonal line show a deviation of 4dB from perfect prediction, which corresponds to the 95% confidence interval given by the measurement accuracy. Furthermore for each plot the resulting correlation between observed and predicted SRT is displayed.

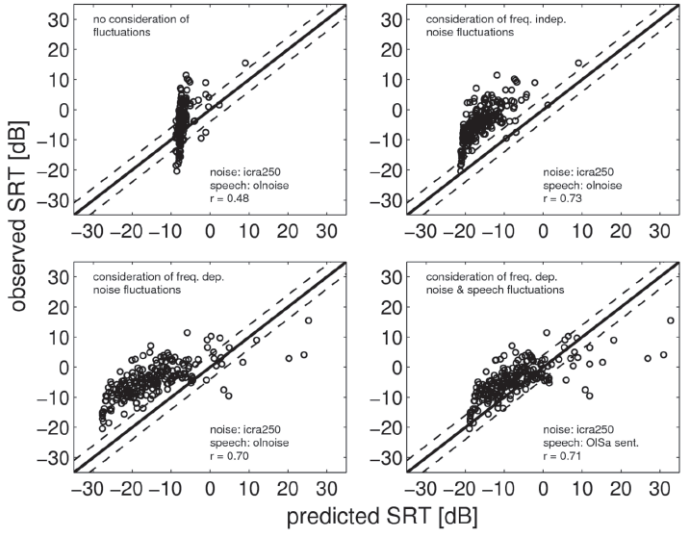

Fig. 3: Observed and predicted Speech reception thresholds (SRT) for four different model versions using fluctuating speech-simulating fluctuating background noise (ICRA 5-250).

The correlation for the standard SII is r=0.48. For the three other model versions the correlation is about r=0.7. This means that the consideration of some temporal information in terms of the frequency independent fluctuations of the noise (Brand and Kollmeier, 2002a) results in a higher correlation between the predicted and the measured SRTs. However, considering further temporal information in terms of the frequency dependent fluctuations of the noise (Rhebergen *et al.* 2005) and the frequency dependent fluctuations of the noise and the speech (extension presented in this study) does not result in a significantly higher correlation.

On the other hand, the consideration of more temporal information yields a closer alignment between predictions and observations even though the correlation does not improve. Although the SII was not designed to predict SRTs in fluctuating noise, it yields good results for some subjects, i.e. about 50% of the predictions are within the 4dB interval. If we consider the frequency independent temporal information (model version based on Brand and Kollmeier, 2002a) the correlation is higher as for the SII (SII: r=0.48, Brand: r= 0.73), but there are less points close to the diagonal, only about 4% of the predictions are within the 4dB interval. If we also take the frequency dependent fluctuations of the noise into account (model version based on Rhebergen *et al.* (2005)) the correlation is slightly smaller than for the version from Brand and Kollmeier, 2002a (Brand: r=0.73, Rhebergen: r=0.70). However, more predictions are close to the diagonal, i.e. about 8% of the predictions are within the 4dB interval. If the fluctuations of the speech signal is accounted for by the extension introduced in this study, the correlation r=0.71 is between the version from Brand and Kollmeier (2002a) and Rhebergen *et al.*, (2005). Even though the predictions for some subjects show a considerable deviation from the observed SRTs, the predictions for most subjects matches the data quite well (about 50% of the predictions are within the 4dB interval). This model version therefore compares favourably to the other model versions that achieve a match of 4% and 8%, respectively. However, this improvement of prediction accuracy is connected with a much higher amount of computational complexity.

Conclusions

As a conclusion for the acoustical level in modelling speech reception, it is save to say that articulation index-based approaches (AI, STI, SII and modifications) appear to work well for threshold-dominated prediction tasks, i. e., for subjects with a mild to moderate hearing loss, for predictions in quiet and the average effect of continuous noise. However, the SRT in stationary noise is only partially predictable for hearing-impaired listeners since the variability among listeners seems to be highly influenced by non-acoustical factors (such as, e. g. sensory effects and cognitive effects). The short-term model extension evaluated here for fluctuating noise seems to work well if spectro-temporal information of both signal and noise is accounted for within the approach. Taken together, modelling the acoustical level appears to be quite successful in assessing the speech information actually present in the acoustical signal. However, since this approach assumes a perfect sensory and cognitive system of the listener, it can only cover individual differences to a certain degree (mostly by accounting for the individual absolute threshold). Hence, the subsequent stages in auditory

processing should be included in the modelling chain to achieve a better understanding of the whole system.

THE SENSORY LAYER: MODEL OF "EFFECTIVE" SIGNAL PROCESSING

A more refined approximation of modelling speech reception should take into account properties of the signal processing in the normal and hearing-impaired auditory system that reflect sensory processes in audition, i.e., the first steps of the physiological transformation of sound into neural activity and the neural representation of sound in the auditory system. Hence, the sensory layer can be thought of as an intermediate stage between the pure acoustical layer and a (perfectly operating) cognitive stage. This sensory layer can therefore be characterized by auditory models of "effective" signal processing that describe the neural transformation from the acoustical signal into some internal neural stage. We assume that the imperfections of the sensory processes involved in human auditory signal processing cause the main limitation in recognizing and discriminating speech sounds. These imperfections should be influenced by auditory signal processing properties that are relevant for human perception of sound, such as, e. g., bandwidth of the "effective" auditory critical bands, compression and adaptation in the auditory system, fine structure versus envelope cues and binaural interaction. The output of the sensory layer is fed into a cognitive layer that exploits the "internal representation" of speech signals in a perfect way by utilizing a-priori knowledge. This layer therefore can be modelled as an "optimal detector" which is assumed to include the whole "world knowledge" of the observer. In a more realistic approach, the cognitive layer can be approximated by a speech recognizer (see below).

One approach to the sensory layer that aims at describing the binaural interaction and binaural noise reduction in normal and hearing-impaired listeners during speech reception tasks was proposed by Beutelmann and Brand (2006). It is based on previous work by vom Hövel (1984) and a similar approach by Zurek (1990). A modification of the equalisation and cancellation (EC-) model of binaural interaction introduced by Durlach (1963) was used as a front end to an SII-type speech intelligibility prediction method.

A more direct way of addressing the sensory component in modelling speech reception was pursued by Holube and Kollmeier (1996) who used an "effective" signal processing model (Dau *et al.*, 1996) of the normal and hearing-impaired listener as a front end to a standard Dynamic-Time Warp (DTW) speech recognizer. By determining the distances between a test utterance and training utterances "on a perceptual scale" (i.e., at the output of the "effective" signal processing model), the utterance with the least distance is taken as the recognized one. The "effective" auditory perception model emplyed (Dau et al,.1996) has been shown to model many different psychoacoustical experiments with different masking conditions as well as modulation detection tasks (Dau *et al.*, 1997).

This approach of combining a perceptual signal processing model (representing the sensory layer) with a DTW speech recognizer (representing the cognitive layer) was

further developed by Jürgens *et al.* (2007). They used for validation the context-free speech database "Oldenburg Logatome Corpus (OLLO)" (Wesker *et al.*, 2005). It contains 70 vowel-consonant-vowel (VCV) and 80 CVC logatomes with the outer phonemes being identical. Each logatome was recorded 18 times by the same speaker spoken in 6 different speech articulation styles: "slow", "normal", "fast", "loud", "quiet" and "questioning". The use of this corpus allows systematical investigations of phoneme recognition rates and confusions. At the same time it avoids that human listeners can use any semantic knowledge for intelligibility.

Human speech recognition (HSR) performance with this speech corpus was measured with 10 clinically normal-hearing subjects. Their age varied between 19 and 37 years. The intelligibility of 150 logatomes was measured in a sound insulated booth at different signal-to-noise-ratios. All recordings were taken from the OLLO database and were spoken by a single German speaker with speech variability "normal".

Jürgens *et al.* combined the Dau *et al.* (1997) model with a standard DTW speech recognizer to mimic the decision process in a closed speech intelligibility test. The level of the template speech waveform is set to 60 dB SPL and both the background ICRA-noise and a hearing threshold simulating noise for normal-hearing listeners is added. The resulting waveform is filtered using a gammatone filterbank (Hohmann, 2002) with 27 frequency channels between 236 Hz and 8 kHz equally spaced on an ERB-scale. The filter outputs are half-wave rectified and low-pass filtered at 1 kHz in a hair cell model. After processing with five consecutive adaptation loops with time constants chosen as in (Holube and Kollmeier, 1996) the signal is again filtered by a modulation filterbank that consists of 4 modulation filters: one low pass at 2.5 Hz and three band passes with center frequencies of 5, 7.5 and 10 Hz and bandwidths of 5 Hz, respectively. The outcome is an "internal representation" (IR) of the time signal. The test signal superposed with a noise waveform is pre-processed in the same way by the perception model. Note that "noise" in this scheme means running ICRA background noise added to a running hearing threshold simulating noise for normal-hearing subjects. All samples of the training vocabulary were equalized to the same length before processing by attaching silence. This was done to rule out a possible discrimination cue due to the individual length of the speech recordings.

The IR of the template and the IR of the test signal are the inputs of the speech recognizer, that calculates the Euclidian distance between the two whereas a DTW algorithm (Sakoe and Chiba, 1978) performs local stretching and compression of the time axes of both IRs in order to achieve a minimal distance. The logatome with the least distance is chosen as the recognized one. The response alternatives given to the model were the same as for HSR.

Two model configurations were realized in this study:

- In configuration A there were 5 IRs per logatome as templates. None of the 5 original recordings was identical to the tested time signal. The logatome that yielded the minimum mean distance of all 5 IRs was chosen as the recognized one.

- Model configuration B contained one IR per logatome as a template whereas the original speech material was identical to that of the test signal. Thus the resulting IRs differ only in the initially added noises.

There are many combinations possible to select speech material from OLLO for performing these model calculations. For these two model configurations the speech recognizing task was calculated 10 times using each time a new combination of speech recordings spoken by the same speaker.

Model predictions and comparison with listening tests

The resulting psychometric functions of the automatic speech recognition (ASR) experiments are shown in Fig. 4. Additionally the fitted psychometric function for normal-hearing subjects from HSR is given as a reference. Configuration A results in a SRT of 1,3 dB calculated from the fitted psychometric function which is more than 13 dB higher than that in HSR. It was assumed that in this model configuration, which closely resembles ASR tasks, 100 % model recognition rate can never be achieved even without background noise. This is due to the inherent speech variability that is still a problem in ASR tasks (Lippmann, 1997). To include this fact a third parameter (the difference between 100 % and the saturation hit rate of the model) was introduced into the fitting routine. With a slope of 5,8 %/dB the reference slope is reproduced quite well.

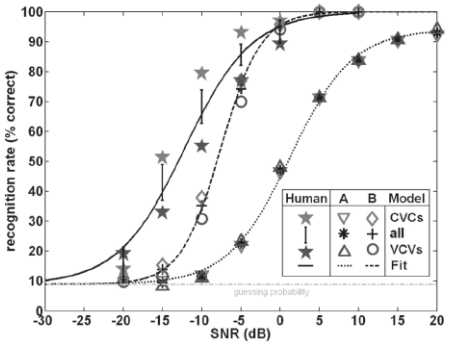

Fig 4: Measured (Human Speech Recognition with normal-hearing subjects, solid curve) and predicted psychometric functions for model configurations A and B derived with utterances of logatomes in ICRA-noise as a function of SNR.

A much better prediction of the normal-hearing psychometric function is achieved with model configuration B. The order of CVC and VCV as well as the upper part of the reference curve is modelled correctly. 100 % recognition rate is reached at 10 dB SNR. The slope (8.9 %/dB) deviates slightly from the reference, the SRT (-7.6 dB) is much closer to human listeners SRT, but still there is a gap of 4.6 dB between them.

Our results show that the psychometric function can only be predicted well if identical speech test and training utterances are used as inputs for the model. This indicates that the variability across speech items is not a crucial factor for understanding speech intelligibility in normal-hearing subjects because they can recognize an

unknown speech waveform that roughly resembles the "expected" template as well as if they would have perfect a priori knowledge of the waveform to expect. The model, on the other hand, is not able to "generalize" the a priori knowledge of a known waveform to a similar utterance. Hence, a model that does not hold the exact speech recording in its training vocabulary performs much worse that a model with perfect a priori information.

Conclusions

As a conclusion for the sensory layer we can state that the prediction of speech reception appears to be quite successful if an "ideal detector" is assumed, i. e., a perfect world knowledge of the word to be expected. In such a configuration, an "effective model" of auditory signal processing seems to predict the availability of speech cues quite well. This is markedly different from the speech intelligibility index-based approach discussed above because speech discrimination is directly predicted from the speech signal without any prior normalisation of the intelligibility function for the respective speech material to be expected. On the other hand, the assumption of a perfect world knowledge (i. e., previous knowledge of the word to be expected as a kind of "Wizard of Oz" experiment) is only a very rough model of the cognitive system and does not take into account any individual differences in cognitive processing abilities. This calls for a better modelling of speech reception including the cognitive level.

THE COGNITIVE LAYER: A "FAIR" MAN-MACHINE COMPARISON

Several approaches exist in the literature to examine the influence of inter-individual cognitive factors on obtained speech reception thresholds for normal and hearing-impaired listeners. The Linköping group (Larsby *et al.*, 2005), for example, could demonstrate a high correlation between speech reception thresholds in noise and cognitive test outcomes, such as tests for assessing the individual working memory and maximum cognitive load by, e. g., performing a dual task memory span experiment. Based on their work, a cognitive test was included in the Hearcom auditory profile which is currently under consideration in a multicenter trial (Dreschler *et al.*, 2007). However, in order to model the cognitive component in a more quantitative way and in order to connect this to models of the acoustical and sensory level (as given above), one will have to exchange the "ideal observer" concept outlined above with a "realistic observer" concept which includes a realistic pattern recognition model and various training procedures to account for priori knowledge in a scalable way. The best currently available pattern recognizers for speech stimuli are highly developed within the field of automatic speech recognition (ASR) so that a model of human speech recognition (HSR) based on elements of automatic speech recognition appears to be a meaningful approach. Since human listeners outperform ASR systems in almost all experiments (Lippmann, 1997), ASR may also profit from auditory feature extraction as proposed in Kleinschmidt, (2003) or by using models of human word recognition (Scharenborg, 2005). In addition, a comparison between HSR and ASR should provide an appropriate basis for advancing such models of human speech recognition. Ideally, such a refined model should not only utilize bottom-up processes (such as, transform-

ing the acoustical input signal into an internal representation which is recognized by a more or less ideal pattern recognizer), but should also incorporate aspects of top-down processing (such as, e. g. using learned patterns and a hypothesis-driven pattern recognition that may be influenced by the individual´s cognitive competence and working memory limitations) in order to model speech recognition in a more adequate way.

As a first step into this direction, a "fair" comparison of human and machine phoneme recognition was achieved by Meyer *et al.* (2007): For similar experimental conditions, the OLLO speech database as described above with non-sense syllables was used for ASR and HSR tests. Hence, human listeners were not able to exploit context knowledge and language models in ASR could be disregarded. This helps to decouple the influence of two major sources of errors in ASR, namely the feature extraction stage and the back-end. Different error patterns of the confusions of phonemes should help to identify sources of errors and to improve ASR feature extraction. It was also investigated whether the information contained in ASR features is sufficient for human listeners to recognize speech. Therefore, feature vectors used internally by the speech recognizer were decoded to acoustic speech tokens. The most common features in ASR (Mel Frequency Cepstral Coefficients / MFCCs) were resynthesized to audible signals which were presented to human test subjects.

Since the calculation of MFCCs results in a loss of information (as, e .g., phase, spectral resolution and fundamental frequency), these signals sound artificial and tinny (like synthesized speech). For example, the speaker's identity or even gender are usually not recognizable. Nevertheless, the resynthesized logatomes are perfectly understandable in the absence of noise. To allow for a valid comparison, the presented recognition scores were obtained with noisy speech (0 dB SNR). By adding noise, redundant information in the speech signal is masked, so that intelligibility is potentially decreased in contrast to an unprocessed signal. The reduction of redundancy might be particularly critical in the presence of speech intrinsic variabilities as, for example, regional dialect.

In order to decode the features to an acoustic speech signal, a linear neural network trained with the OLLO training set is used to construct the spectral envelope from the cepstral coefficients. Additional information such as voicing or fundamental frequency fg is not used for the calculation, since this would give human listeners an unfair advantage over ASR. Hence, an artificial excitation signal has to be used. Pilot experiments showed that intelligibility is highest when a pulse train with fg = 130 Hz is used as excitation signal (instead of noise or a mixed noise-pulse signal). In a final step, the spectral envelope and the artificial excitation signal are combined. This algorithm was kindly supplied by the Katholieke Universiteit Leuven (Demuynck, 2004).

During the extensive HSR experiments, the original signals in speech-shaped noise at an SNR of -10 dB were also presented as a reference condition. Five normal-hearing listeners participated in the tests. Their task was to identify the middle phoneme in OLLO VCV and CVC utterances. The recognition rates were compared to ASR results obtained with a standard HMM recognizer with MFCC feature extraction (see Fig. 5).

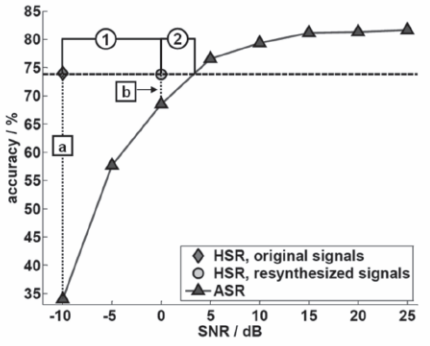

Fig. 5: HSR and ASR phoneme recognition scores over SNR. The large gap between HSR and ASR (a) is reduced if human listeners are provided with only the information contained in ASR features (b). The gap expressed in terms of the SNR may be attributed to either the sensory or cognitive differences between HSR and ASR (labels 1 and 2, respectively).

A direct comparison of ASR and HSR performance shows that human speech recognition is superior to the ASR system under equal conditions, as displayed in Fig. 5 (from Meyer *et al.*, 2007). The total HSR and ASR accuracies at an SNR of -10 dB are 74.0% and 34.0%, respectively, which corresponds to a relative increase of the word error rate (WER) of 154% (dotted line a). The gap narrows if the information for human listeners is limited to the information content of MFCCs: For resynthesized signals at 0 dB SNR, the recognition score is 73.8% and the corresponding ASR accuracy is 68.5%, resulting in a WER increase of 20.0% (dotted line b).The SNRs for both HSR conditions were chosen so that average recognition rates are similar. The choice of SNRs was based on the presentation of only few test lists to one human listener and proved to be reasonable for other test subjects as well, as the overall accuracies are very close to each other: The average scores were 73.8% (resynthesized signals) and 74.0% (original signals), respectively. Therefore, the information loss caused by MFCCs can be expressed in terms of the signal-to-noise ratio, i.e. the SNR of resynthesized signals has to be 10 dB higher in order to obtain similar recognition performance (label 1 in Fig. 5). The difference between HSR with resynthesized signals and ASR is 3.2 dB (label 2 in Fig. 5). The results can be interpreted as follows: The total gap between human and automatic speech recognition in terms of SRT amounts to 13.2 dB (i. e. ASR at +3.2 dB yields the same recognition score as HSR at -10 dB SNR). This gap can be separated into a "sensory part", i. e. the gap between HSR for natural speech and for re-synthesized speech (i.e., label 1) which amounts to 10 dB. It is due to non-ideal representation of the speech signal as the input pattern for the speech pattern recognition model. The remaining gap of 3.2 dB between HSR for resynthesized speech and ASR (i.e., label 2) can be interpreted as the "cognitive" gap, i. e., the advantage of human "top-down"-processing over the statistical-model-based pattern recognition in the ASR. Even though the HMM speech recognizer employed here is only a poor model of the human cognitive system in recognising speech, this comparison still helps to quantitatively assess the effect of cognition for speech recognition in noise. Inter-

estingly, the 3.2-dB gap found here is in the same order of magnitude as the difference between native and non-native listeners found in SRT measurements with sentences (for example Warzybok *et al.*, 2007).

As a conclusion for the cognitive level, we can say that no promising "Ansatz" exists yet to adequately model the cognitive level in speech recognition. Hence, more work will have to be invested to achieve a satisfactory, complete model that will eventually also include individual differences in cognitive processing for the prediction of speech reception thresholds. However, the comparison between the perceptual, information-driven approach (bottom-up) and the world knowledge- and hypothesis-driven approach (top-down) pursued here appears to be a reasonable first step.

GENERAL CONCLUSION

As a very rough estimate of how much we already understand about speech recognition with our model approaches at different layers, we could state:

- On the acoustical layer using prediction methods that are based on various speech-to-noise-ratio measures (such as the articulation index and speech intelligibility index plus derivatives) we already achieve a very high prediction quality. The methods are already very elaborated for a long-term speech intelligibility predictions, for (nearly) linear systems and for auditory-threshold-dominated speech perception tasks (i. e., speech perception in quiet for hearing-impaired listeners). A short-term extention appears promising if both the time-varying speech and noise spectrum is adequately accounted for.
- At the sensory layer, the models range from binaural interaction models combined with SII-based speech intelligibility prediction up to models with a microscopic analysis of auditory processing involved in speech perception. Good progress has been made in modelling normal-hearing to moderate hearing- impaired listeners in (nearly) stationary noise conditions and assuming a "perfect" world knowledge. However, a wide range of interesting questions still have to be solved, such as, e.g. an appropriate characterization of the effect of suprathreshold processing deficits in hearing-impared listeners.
- With respect to the cognitive layer, the approach for a "fair" comparison between human and machine speech recognition seems to be a first reasonable approach, especially if the inclusion of a different degree of "world knowledge" and training is accounted for. However, no quantitative model is yet achieved that would relate e. g. the cognitive load and processing costs involved in cognitive processing to the acoustical and the perceptual level. Hence, considerable amount of work still has to be invested here.

ACKNOWLEDGEMENT

Supported by BMBF and research ministry of lower Saxony (Kompetenzzentrum HörTech) and the CEC (project Hearcom). We also thank all members of the Medical

Physics group, the HörTech and the Hearcom consortium for their cooperation as well as the subjects for their patience.

REFERENCES

ANSI (**1997**). "Methods for Calculation of the Speech Intelligibility Index", American National Standard S3.5-1997.

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. 2006. **120**: p. 331-42.

Brand, T., and Kollmeier B. (**2002b**). "Efficient adaptive procedures for threshold and concurrent slope estimates for psychophysics and speech intelligibility tests," J. Acoust. Soc. Am. 2002. **111**(6): p. 2801-2810.

Brand, T., and Kollmeier, B. (**2002a**). "Vorhersage der Sprachverständlichkeit in Ruhe und im Störgeräusch aufgrund des Reintonaudiogramms," DGA 2002.

Dau, T. (**1997**). "Modeling auditory processing of amplitude modulation," J. Acoust. Soc. Am. 1997. **101**: p. 3061 (A).

Dau, T., Püschel, D., and Kohlrausch A. (**1996**). "A quantitative model of the "effective" signal processing in the auditory system: I. Model structure," 1996. **99**: p. 3615-3622.

Dau, T., Kollmeier, B., and Kohlrausch, A. (**1997**). "Modeling auditory processing of amplitude modulation: I. Detection and masking with narrow band carrier," J. Acoust. Soc. Am., **102**, 2892-2905.

Demuynck, K., Garcia, O., and Dirk Van Compernolle, (**2004**): "Synthesizing Speech from Speech Recognition Parameters," In Proc. ICSLP 2004, vol. II, 945–948.

Dreschler, W. A., (**2001**). "ICRA Noises: Artificial Noise Signals with Speech-like Spectral and Temporal Properties for Hearing Instrument Assessment," Audiology. **40**, 148–157.

Dreschler, W. A., van Esch, T. E. M., and Jeroen Sol, J. (**2007**). "Diagnosis of impaired speech perception by means of the Auditory Profile," this volume.

Durlach, N. I. (**1963**). "Equalization and cancellation theory of binaural masking-level differences", J. Acoust. Soc. Am. **35**(8), p. 1206–1218.

Fletcher, H., Galt (**1950**). "The Perception of Speech and Its Relation to Telephony," J. Acoust. Soc. Am. 1950 **22**(2), 89-151.

Hohmann, V. (**2002**). "Frequency analysis and synthesis using a Gammatone filterbank," Acta acustica / Acustica, 2002. **88**(3). 433-442.

Holube, I., and B. Kollmeier (**1996**). "Speech intelligibility prediction in hearing-impaired listeners based on a psychoacoustically motivated perception model," J. Acoust. Soc. Am., 1996. **100**(3). 1703-16.

Houtgast, T., Steeneken, H. J. M. (**1985**). "A review of the MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria," J. Acoust. Soc. Am. **77**(3). 1069-1077.

Jürgens, T., Brand, T., and Kollmeier, B. (**2007**). "Modelling the Human-Machine Gap in Speech Reception: Microscopic Speech Intelligibility Prediction for Normal-Hearing Subjects with an Auditory Model," In Proc. Interspeech 2007, Antwerpen.

Kleinschmidt, M. (**2003**). "Localized spectro-temporal features for automatic speech recognition," Proc. Eurospeech/Interspeech, Geneva, 2003.

Kollmeier, B. (**1990**). "Meßmethodik, Modellierung und Verbesserung der Verständlichkeit von Sprache," Habilitationsschrift, Universität Göttingen

Larsby, B., Hällgren, M., Lyxell, B., Arlinger, S. (**2005**). "Cognitive performance and perceived effort in speech processing tasks: effects of different noise backgrounds in normal-hearing and hearing-impaired subjects," Int. J. Audiol. **44**(3), 131-143.

Lippmann, R.P. (**1997**). "Speech recognition by machines and humans," Speech Communication **22** (1), 1-15, 1997.

Meyer, B., Wächter, M., Brand, T. and Kollmeier, B. (**2007**): "Phoneme confusions in human and automatic speech recognition," In Proc. Interspeech 2007, Antwerpen, Belgium, 2007.

Meyer, R., Brand, T. (**2007**). "Prediction of speech intelligibility in fluctuating noise," in: EFAS/DGA 2007, Heidelberg (in press).

Pavlovic, C. V. (**1984**). "Use of the articulation index for assessing residual auditory function in listeners with sensorineural hearing impairment," J. Acoust. Soc. Am. **75**(4), 1253–1258.

Payton, K.L., Braida, L.D. (**1999**). "A method to determine the speech transmission index from speech waveforms," J. Acoust. Soc. Am. **106**(6), 3637-3648.

Plomp, R. (**1986**). "A Signal-to-Noise Ratio Model for the Speech-Reception Threshold of the Hearing Impaired," J.Sp.Hear. Res. **29**, 146-154.

Rankovic, C. M. (**1997**). "Prediction of Speech Reception by Listeners With Sensorineural Hearing Loss," in: Jestaeadt, W. (ed.): Modeling Sensorineural Hearing Loss, Lawrence Earlbaum Associates, Mahwah, N.J.

Rhebergen, K., and Versfeld, N. (**2005**). "A Speech Intelligibility Index-based approach to predict the speech reception threshold for sentences in fluctuating noise for normal-hearing listeners," J. Acoust. Soc. Am. **117**, 2181-92.

Sakoe, H., and S. Chiba (**1978**). "Dynamic Programming Algorithm Optimization for Spoken Word Recognition," IEEE Transactions on Acoustics, Speech, and Signal Processing, ASSP-26(1), 43-49.

Scharenborg, O. (**2005**). "Narrowing the gap between automatic and human word recognition," Ph.D. thesis, Radboud University Nijmegen, September 16th, 2005.

Sroka, J. J., and Braida, L. D. (**2005**). "Human and Machine Consonant Recognition," Speech Communication **45** (401-423), 2005.

Vom Hövel, H. (**1984**). "Zur Bedeutung der Übertragungseigenschaften des Außenohrs sowie des binauralen Hörsystems bei Gestörter Sprachübertragung," Dissertation, Fakultät für Elektrotechnik, RTWH Aachen.

Wagener, K., Brand T., Kollmeier, B. (**2006**). "The role of silent intervals for sentence intelligibility in fluctuating noise in hearing-impaired listeners," Int. J. Audiol. **45**, 26-3.

Wagener, K., Brand, T., Kühnel, V., and Kollmeier, B. (**1999**). „Entwicklung und Evaluation eines Satztestes für die deutsche Sprache I-III: Design, Optimierung und Evaluation des Oldenburger Satztestes," Zeitschrift für Audiologie **38**(1-3)

Wagener, K., and Brand, T. (**2005**). "Sentence intelligibility in noise for listeners with normal hearing and hearing impairment: influence of measurement procedure and

masking parameters," Int. J. of Audiol. **44**(3), p. 144-156.

Wagener, K., Brand, T., and Kollmeier, B. (**2006**). "The role of silent intervals for sentence intelligibility in fluctuating noise in hearing--impaired listeners," Int. J. of Audiol. **45**(1), 26-33.

Wagener, K. C., Brand, T., and Kollmeier, B (**2007**). "International cross-validation of sentence intelligibility tests", EFAS - Meeting 2007, Heidelberg (in press).

Warzybok, A., Wagener, K.C., Brand, T. (**2007**). "Intelligibility of German digit triplets for non-native German listeners," EFAS - Meeting 2007, Heidelberg (in press)

Wesker, T., Meyer, B., Wagener, K., Anemüller, J., Mertins, A., and Kollmeier, B. (**2005**). "Oldenburg Logatome Speech Corpus (OLLO) for Speech Recognition Experiments with Humans and Machines," In Proc. Interspeech 2005, Lisbon, Portugal, 1273-1276.

Zurek, P. M. (**1990**). "Binaural advantages and directional effects in speech intelligibility," in Acoustical Factors Affecting Hearing Aid Performance, 2nd ed., edited by G. A. Studebaker and I. Hockberg, Allyn and Bacon, London, Chap. 15, 255–276.