Consonant recognition of listeners with hearing impairment and comparison to predictions using an auditory model

T. Jürgens, T. Brand and B. Kollmeier

Carl-von-Ossietzky Universität Oldenburg, Germany, Email: tim.juergens@uni-oldenburg.de

Introduction

Listeners with sensorineural hearing impairment show a substantially decreased speech recognition performance compared to normal-hearing listeners. For listeners with high-frequency hearing loss, particularly consonant recognition is degraded in noisy environments but also in quiet. In a recent study [1] it was shown that consonant recognition of normal-hearing listeners in noise can be predicted well by an auditory model combined with a speech recognizer. To test the hypothesis that audibility is not the only factor influencing consonant recognition performance of hearing-impaired listeners in quiet conditions, the model of [1] is extended to hearing-impairment. Different model variations implementing hearing-impairment are tested regarding either audibility only, or additionally a changed dynamic compression. Model results are compared to observed consonant recognition rates of normal-hearing and hearing-impaired listeners.

Method

Subjects

Ten normal-hearing subjects (aged 21-38 years) and four hearing-impaired subjects (aged 52-67 years) were employed. Audiometric thresholds were 10 dB HL maximum at any audiometric frequency for the normal-hearing listeners. The audiometric thresholds of the hearing-impaired listeners are shown in figure 1.

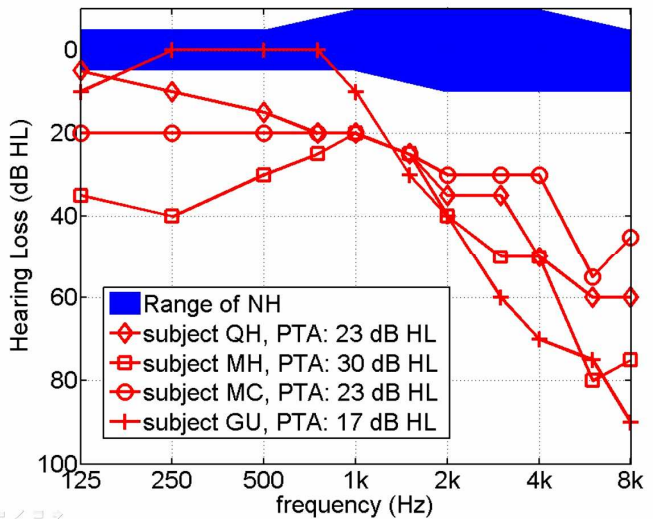

Figure 1: Pure tone hearing thresholds (air conduction) of four hearing-impaired listeners (red symbols and solid lines) and data range of ten normal-hearing (NH) listeners (blue area). Pure-Tone-Average (PTA) values of the hearing-impaired listeners are indicated in the legend.

The bone-air gap did not exceed 10 dB indicating sensorineural hearing loss. Hearing loss was symmetric at both ears ($\leq$ 15 dB difference between right and left ear) but only the better ear was chosen for the speech tests.

Speech tests

Consonant recognition was measured by presenting nonsense vowel-consonant-vowel (VCV) utterances from the OLdenburg LOgatom (OLLO) speech corpus [2]. The OLLO corpus is freely available at http://sirius.physik.uni-oldenburg.de. Levels ranged from 5 to 25 dB SPL for normal-hearing and 25 to 60 dB SPL for hearing-impaired listeners. A closed-set paradigm was used providing response alternatives that differ only in the consonants: /p/, /t/, /k/, /b/, /d/, /g/, /s/, /f/, /v/, /n/, /m/, /ʃ/, /ts/, /l/. Preceding and subsequent vowels were kept the same for all response alternatives. The consonants were embedded in the vowels /a/, /ɛ/, /ɪ/, /ɔ/, or /ʊ/. The speech waveforms were presented monaurally in a sound-insulated booth via headphones Sennheiser HDA 200 that were digitally free-field equalized.

Model structure

Figure 2 shows the structure of the CASP (Computational Auditory Signal-processing and Perception) model by Jepsen et al. [3] that was used for the "auditory" preprocessing of the speech waveforms. It computes from an audio signal an "internal representation" by using the following stages: The first stage is a digital filter modelling the transformation through the outer- and middle ears. The most characteristic feature of the CASP model is the second stage that is a Dual-Resonance-NonLinear (DRNL) filterbank that mimics the complex nonlinear Input/Output-(I/O-) characteristic of the Basilar-Membrane (BM). The DRNL filterbank consists of two processing paths whose outputs are added: the upper linear processing path is a broad bandpass gammatone filter (GFB) and the lower, nonlinear path is a very sharp GFB at the same centre frequency. For low-level on-frequency signals the lower path amplifies the signal instantaneously. High-level on-frequency signals are not amplified resulting in a compressive I/O-characteristic for on-frequency signals. Remote off-frequency signals are processed linearly. The mechanical-to-neural transduction is then modelled by a haircell-model, a squaring expansion, an adaptation stage and a modulation filterbank. The adaptation stage performs a logarithmic dynamic compression for stationary signals and emphasizes on- and offsets of non-stationary signals. Modelling speech recognition is performed as proposed by Holube and Kollmeier [4]: For each possible response alternative an internal representation is computed using CASP. The speech waveform to recognize is processed in

the same way. A Dynamic-Time-Warp [5] speech recognizer is used to compute the distance between the internal representations of a response alternative and the speech waveform to recognize. The response alternative yielding the smallest distance to the speech waveform to recognize is taken as the "recognized" one.

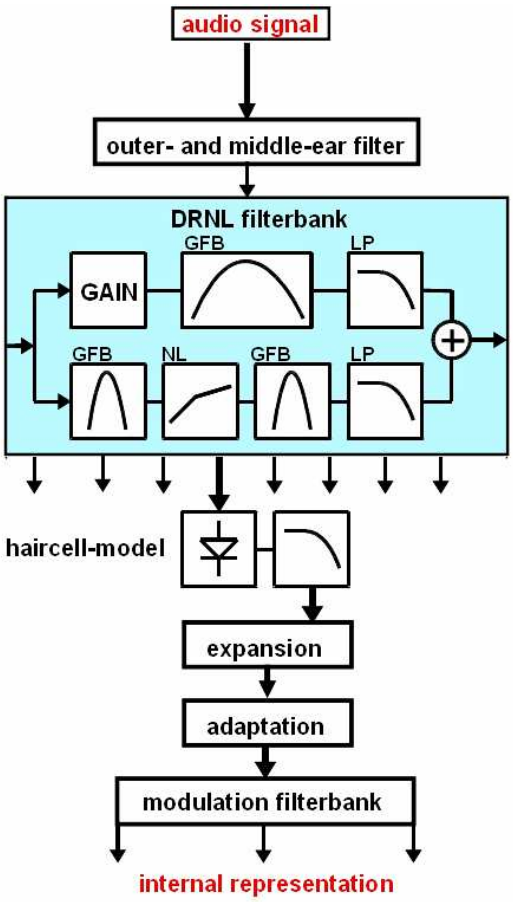

Figure 2: Sketch of the auditory preprocessing stage taken from [3]. From an audio signal an "internal representation" is calculated. See text for a further description of the model.

Model variations

To investigate the influence of different changes in the model reflecting different pathologic changes in the impaired auditory system, three model variations were used:

1. A hearing threshold simulating noise is added to the speech waveform before entering the CASP model. Running steady-state noise was used that was spectrally shaped to the individual subjects' audiogram data. This model variation is called **"external noise"**.
2. Additionally to the external noise from model variation 1 the compressive properties of the DRNL filterbank are changed in a way that less amplification to low-level signals is provided in the nonlinear processing path. This results in a more linearized behaviour of the DRNL filterbank to on-frequency signals. This model variation is called **"external noise + attenuation"**.
3. An internal hearing threshold simulating noise is added to the output of each DRNL filterbank channel, respectively. This internal noise was calibrated by measuring the output of the DRNL filterbank channels when only the external noise for normal-hearing listeners is processed by CASP. Additionally, the compressive properties of the DRNL filterbank were changed as in model variation 2. This model variation is called **"internal noise + attenuation"**.

The first model variation can be interpreted as that audibility only is taken into account as factor contributing to the poorer speech recognition performance of hearing-impaired listeners. The second variation regards the changed I/O-characteristics of the sensorineurally hearing impaired system as was observed by Plack et al. [6], for instance, additionally to the reduced audibility. The attenuation in the nonlinear processing path is directly attributed to the amount of hearing loss due to the loss of outer hair cells. In the third model variation the internal noise can be interpreted as spontaneous firing of inner hair cells that is assumed to be the same for both normal-hearing and hearing-impaired listeners if no additional inner hair cell loss is present. Hearing Loss is split in a part attributed to the loss of outer hair cells and resulting in a "linearization" of the BM I/O-function and a part attributed to the loss of inner hair cells resulting in an increased internal noise. The amount of outer hair cell loss is assumed to be 80% of the hearing loss from audiogram data, at maximum 40 dB, whereas the amount of inner hair cells is the remainder.

Test conditions

Calculations with the speech recognition model as well as measurements with human listeners were performed under highly similar conditions: The same speech waveforms from OLLO were used. The same response alternatives were offered to both human listeners and the model.

Results

Normal-hearing listeners

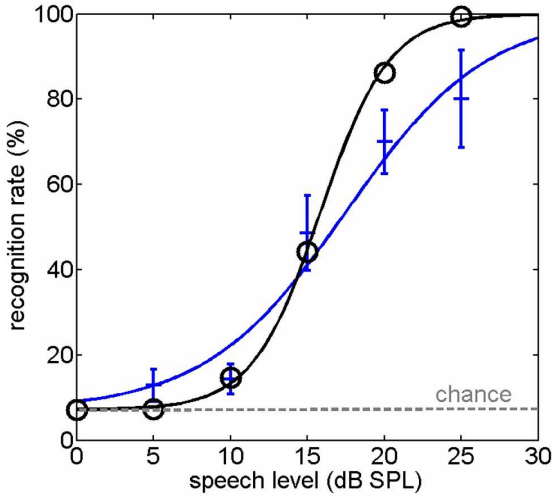

Figure 3: Psychometric function of consonant recognition in quiet condition for ten normal-hearing subjects (blue crosses and blue solid line) and model predictions (black circles and solid black line). Blue error bars denote the inter-individual standard deviation across subjects.

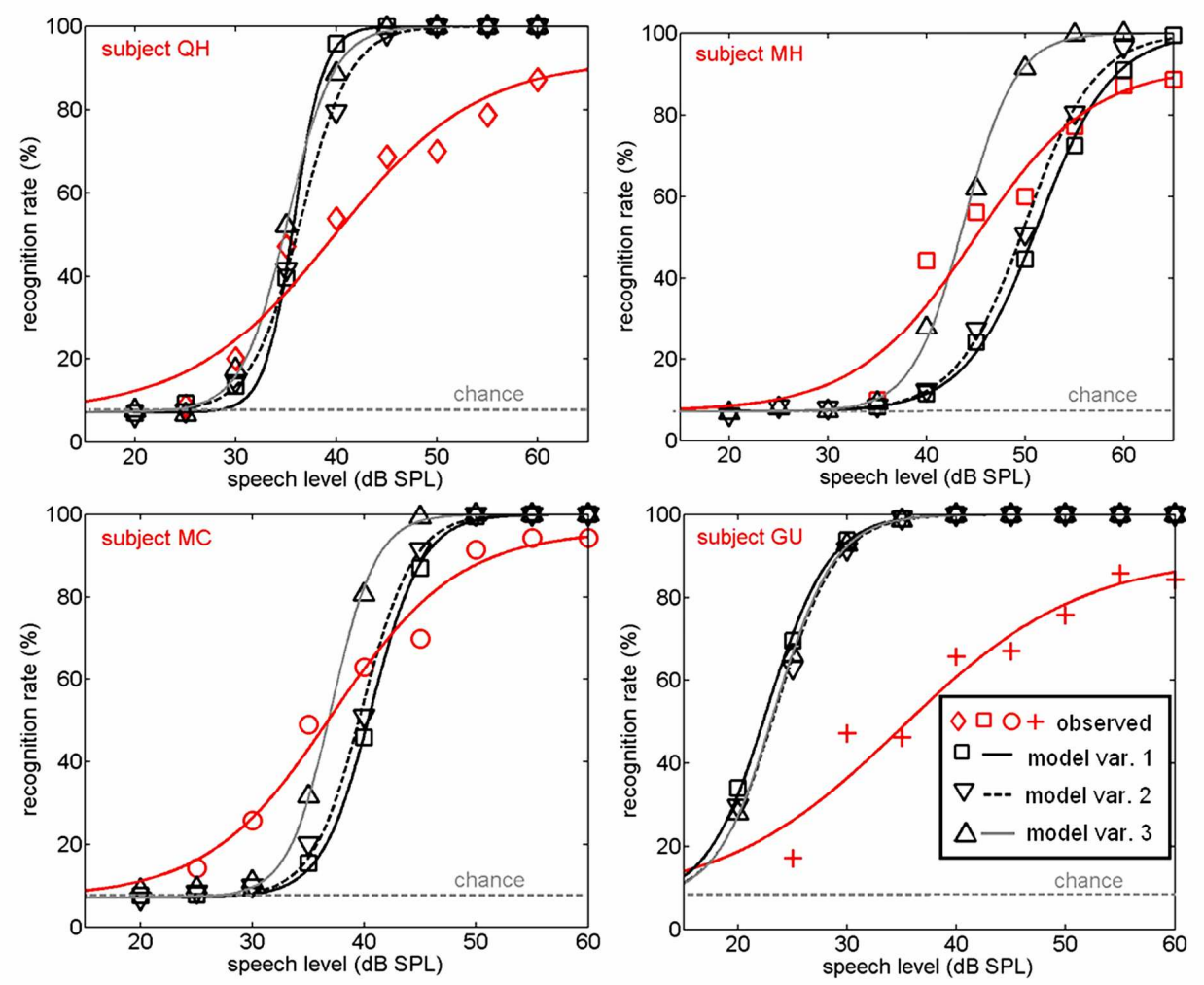

Figure 4: Psychometric functions of consonant recognition in quiet condition for four hearing-impaired subjects (red symbols and lines) and model predictions using model variation 1 ("external noise", squares and black solid line), variation 2 ("external noise + attenuation", downward triangles and dashed line), and variation 3 ("internal noise + attenuation", upward triangles and grey solid line).

Figure 3 shows the psychometric function, i.e. recognition rates in percent correct versus absolute speech level, of consonant recognition by normal-hearing listeners (blue crosses) averaged across the ten subjects. Additionally, the inter-individual standard deviations are plotted as error bars. The data is fitted using a logistic function assuming chance (7.1% recognition rate) at low levels and 100% recognition rate at very high levels (blue solid line). Model predictions are plotted in black using the "external noise" model variation. The results show that the Speech-Reception-Threshold (SRT), i.e. the speech level at 54% recognition rate is predicted with an accuracy of 1-2 dB (cf. table 1). However, the slope of the modelled psychometric function is much steeper than observed by the normal-hearing listeners. This is mainly determined by the two data points at 20 and 25 dB SPL where the model overestimates normal-hearing listeners' performance. Note, that the two other model variations are not plotted in figure 3 because firstly, they provide the same model performance as model variation 1. Secondly, it is not reasonable to change the compressive properties of the DRNL filterbank for normal-hearing listeners because the parameters are tuned to reflect normal-hearing listeners' I/O-function.

Hearing-impaired listeners

The four plots of figure 4 show individual consonant recognition rates for the four hearing-impaired subjects (red symbols) and fitted psychometric functions (red solid lines). Model predictions are additionally plotted using the "external noise" model variation (squares and solid black line), the "external noise + attenuation" variation (downward triangles and dashed black line), and the "internal noise + attenuation" variation (upward triangles and solid grey line). Comparing the model variations among each other shows that the performances of the "external noise" and the "external noise + attenuation" variation are always very close to one another. For subjects with low pure-tone-average (PTA, cf. figure 1) also the "internal noise + attenuation" variation shows the same performance whereas for high PTA the "internal noise" variation shows higher recognition rates. Table 1 shows the observed and predicted SRTs using the different model variations. For the subjects QH, MH, and MC the "internal noise + attenuation" variation shows the best SRT predictions. The accuracy of predicting the SRT is slightly poorer (3-4 dB) than for the normal-hearing listeners. The two "external noise" and "external noise + attenuation" variations show poorer results

for the prediction of hearing-impaired listeners' performance (accuracy about 5 dB). All three model variations overestimate strongly the performance of subject GU (about 12 dB difference in SRT). At large, the predicted psychometric functions of all model variations are much steeper than the observed psychometric functions.

	observed SRT (dB)	SRT of model var.1 (dB)	SRT of model var.2 (dB)	SRT of model var.3 (dB)
normal-hearing	17.5	15.9	-	-
QH	39.7	36.3	36.4	35.1
MH	44.8	51.3	50.0	43.8
MC	37.2	40.7	39.9	37.1
GU	35.5	22.7	23.5	23.4

Table 1: Observed and predicted Speech-Reception-Thresholds (SRTs) of ten normal-hearing listeners (averaged) and four hearing-impaired listeners (individual data) using different model variations. See text for further description of the different model variations.

Discussion

The surprisingly high accuracy at predicting the SRT in quiet for normal-hearing listeners (1-2 dB) shows that the threshold for pure tones is the main factor governing the threshold value for speech in the healthy auditory system. However, the poor prediction of the consonant recognition scores at 20 and 25 dB SPL shows that the steady-state running noise is not sufficient for modelling the details of the speech perception process above SRT. Holube and Kollmeier [4], for instance, assumed a fluctuating hearing threshold noise as a more elaborated approach for modelling speech recognition in quiet.

The nearly identical performances of the "external noise" and "external noise + attenuation" variation of the model for four hearing-impaired subjects show that a change in the compressive properties of the model does not affect modelling consonant recognition in quiet condition. This gives support to the hypothesis that also for the hearing-impaired, audibility is the most important factor for limiting consonant recognition in quiet. Nevertheless, the predictions of the performance of hearing-impaired subjects are slightly poorer using one of these two model variations than using the "internal noise + attenuation" model variation (accuracy of 3-4 dB). This model variation regards hearing loss due to outer- and inner hair cell loss separately. The poor predictions of subject GU may be due to the very steep slope of the audiometric thresholds at high frequencies (about 25 dB/oct.). The also severe hearing loss of >80 dB HL at high frequencies may be an indication of "dead regions" that could give an additional distortion to the speech signal. This was not regarded by the model.

Conclusions

- The prediction of the SRT of consonant recognition for normal-hearing listeners in quiet is achieved by the combination of auditory model and speech recognizer with an accuracy of 1-2 dB.
- Changing the compressive properties of the auditory model to include the suprathreshold factor "reduced dynamic compression" does not affect the predictions in quiet when using an external hearing threshold simulating noise.
- The best SRT predictions for hearing-impaired listeners in quiet are achieved by an internal hearing threshold noise additionally to changed compressive properties.
- There are some observed results (steepness of the psychometric function, performance of one hearing-impaired listener) that could not be predicted using this model which may be attributed to additional suprathreshold factors.

Acknowledgements

The authors would like to thank the SFB TRR 31 „The active auditory system" for funding the research reported in this paper.

References

[1] Jürgens, T., T. Brand, and B. Kollmeier, *Modelling the human-machine gap in speech reception: microscopic speech intelligibility prediction for normal-hearing subjects with an auditory Model*, in *Interspeech 2007*. Antwerp, Belgium. p. 410-413.

[2] Wesker, T., Meyer, B., Wagener, K., Anemüller, J., Mertins, A., and B. Kollmeier, *Oldenburg logatome speech corpus (OLLO) for speech recognition experiments with humans and machines*, in *Interspeech 2005*. Lisboa, Portugal. p. 1273-1276.

[3] Jepsen, M.L., S.D. Ewert, and T. Dau, *A computational model of human auditory signal processing and perception.* J. Acoust. Soc. Am., 2008. **124**(1): p. 422-438.

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

[5] Sakoe, H. and S. Chiba, *Dynamic programming algorithm optimization for spoken word recognition.* IEEE Transactions on Acoustics, Speech, and Signal Processing, 1978. **ASSP-26**(1): p. 43-49.

[6] Plack, C.J., V. Drga, and E.A. Lopez-Poveda, *Inferred basilar-membrane response functions for listeners with mild to moderate sensorineural hearing loss.* J Acoust Soc Am, 2004. **115**(4): p. 1684-95.