Modeling a damaged cochlea: beyond non-speech psychophysics

Morten L. Jepsen[1,2], Oded Ghitza[2], and Torsten Dau[1]

1 Centre for Applied Hearing Research, Technical University of Denmark, Lyngby, Denmark

2 Hearing Research Center, Boston University, Boston, MA, USA

It has long been recognized that audiograms provide only a limited view of hearing impairment; exploiting "beyond audiogram" psychophysics was only recently reported. Here, we explore a method in which speech test stimuli are used as integral part of the modelling process of damaged cochleae. In a preliminary phase, non-speech psychophysical data were collected in individual hearing-impaired listeners. Following a procedure described previously, these data were used to adjust parameters of a peripheral auditory model, aiming at simulating the individual listeners' hearing impairment. In the second phase, the same individuals were tested in a speech task - a diagnostic rhyme test (DRT). Monosyllabic words, organized as minimal pairs, were synthesized such that their acoustic waveforms only differed in the initial diphone's segments. These stimuli were processed by the model obtained in the preliminary phase; the resulting representation was analyzed by a machine mimicking the DRT paradigm, generating acoustic-phonetic error patterns. An important feature of the DRT framework is the separation of errors originated by the front-end from those originated by the back-end. In comparing machine to human, some error patterns were accounted for by the model, indicating that there is a relation between speech and non-speech psychophysics.

INTRODUCTION

Models of the auditory periphery aim at appropriately describing a perceptually relevant "internal representation" (IR) of the incoming sound in the auditory system. To investigate whether the simulated IR provides a good match to the "real" internal representation, it is crucial that front-end and back-end processing are clearly separated (Ghitza, 1993), for example in a speech recognition system. Ghitza (1993) suggested that the cognitive component involved in consonant recognition could be minimized by using the diagnostic rhyme test (DRT, Voiers, 1983), because this test represents a simple binary discrimination task. Messing *et al*. (2009) proposed a modelling framework to predict confusion patterns from DRT in NH listeners. Part of their focus was on separating the peripheral model (front-end) and the detector stage (back-end). They used synthesized DRT diphones which allowed them to use of a simple back-end which measures a perceptual distance to templates in order to keep the back-end errors at a minimum.

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

In the present study, the goal was to investigate how degraded auditory processing, due to cochlear damage, affects speech perception in individual listeners. This was done using the computational auditory signal-processing and perception (CASP) model (Jepsen *et al*., 2008; Jepsen and Dau, 2010) as the front-end and the synthesized DRT and detector of Messing *et al*. (2009) as the back-end. The separation of front-end and back-end processing was of great conceptual importance here, since predicted errors could thus be uniquely associated with the front-end processing and could provide a measure of how well the CASP model, including the simulation of cochlear hearing loss, can describe the actual IR of speech sounds in individual hearing-impaired listeners. For the simulation of hearing impairment, the front-end was fitted to the individual listeners due to non-speech psychophysics following the method suggested in Jepsen and Dau (2010). The peripheral model then remained unchanged for the simulations of DRT error patterns. If the model was able to predict individual error patterns in the DRT task based on individual fits to the non-speech tasks, then degraded performance in speech tasks could be associated with limitations in basic auditory processing in the individuals.

EXPERIMENTAL METHODS

Listeners

Three listeners with mild-to-moderate sensorineural hearing loss participated in this study. Listeners S1, S2 and S3 were 21, 45 and 27 years old, respectively. All these had a hearing loss since their early childhood. S1 and S2 were regular users of hearing aids while S3 did not use hearing aids. Only one ear of each listener was measured in the speech and non-speech tasks. The audiograms of the measured ears are shown in Fig. 1 (open symbols).

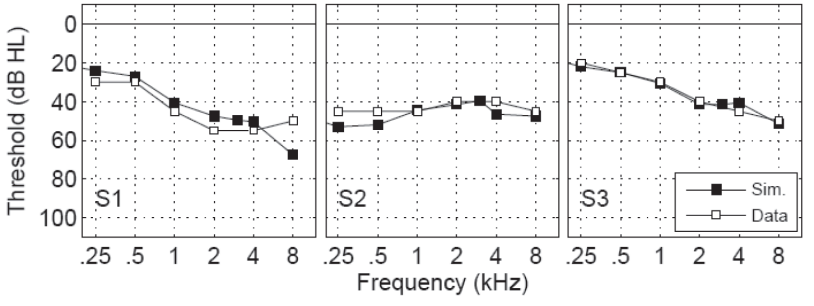


Fig. 1: Audiograms of the measured ears of the three HI listeners. Pure-tone thresholds are plotted in dB hearing level (HL). Open symbols indicate measured thresholds, while filled symbols indicate simulated thresholds by the corresponding models.

Temporal masking curves (TMC)

In the TMC experiment, forward masking of a fixed-level brief tone was measured as a function of signal-masker interval. The probe signal was a pure tone with a

duration of 20 ms, which was Hanning windowed over its entire duration. The signal frequency (f_{sig}) was either 1 or 4 kHz. The signal was presented at 10 dB sensation level (SL). The masker was also a pure tone with a duration of 200 ms and 5-ms raised-cosine on- and off ramps were applied. The masker frequency (f_m) was equal to f_{sig} (on-frequency condition) or 0.6 f_{sig} (off-frequency condition). The masker-signal interval was 2, 5, 10 ms and additional 10-ms increments until the subject reported (in pilot runs) that the masker level became uncomfortably loud. The masker level was adjusted by a adaptive procedure to reach masked signal threshold. A 3-interval 3-alternative forced choice paradigm in connection with a 1-down/2-up rule were applied. The reported thresholds reflect the 70.7% point on the psychometric function and represent the mean of at least three measured thresholds. The step size was varied adaptively, starting at 8 dB and ending at 1 dB, and thresholds were an average of the levels at the last eight reversals with the final step size.

The diagnostic rhyme test (DRT)

The DRT of Voiers (1983) uses 192 minimal pair diphones and was designed to cover six acoustic-phonetic dimensions: voicing (VC), nasality (NS), sustention (ST), sibilation (SI), graveness (GV) and compactness (CM). The diphones were synthesized by a text-to-speech system, such that the acoustic waveforms only differed in the initial consonant. The noise was generated by passing a Gaussian noise through a filter with a speech-shaped transfer function. The speech was presented at a constant *rms* level of 70 dB SPL, and in background noise at a SNR of 0 and 10 dB. Eight repetitions of the 192 diphones were presented in blocks of 128 in random order. The DRT is a binary consonant discrimination task realized in a 1-interval 2-alternative forced-choice method. Examples of minimal pairs are given in (Voiers, 1983; Ghitza, 1993).

MODELING SPEECH PERCEPTION

The front end

A schematic structure of the front-end, the CASP model (Jepsen *et al*., 2008) is shown in Fig. 2. The acoustic stimuli are first processed by the outer- and middle ear filters, followed by the dual-resonance nonlinear (DRNL) filterbank simulating BM processing. The processing of the subsequent stages is carried out in parallel in the frequency channels. Inner hair-cell transduction is modelled roughly by half-wave rectification followed by a first-order lowpass filter with a cut-off frequency at 1 kHz. The expansion stage transforms the output of the IHC stage into an intensity-like representation by applying a squaring expansion. The adaptation stage simulates dynamic changes in the gain of the system in response to changes in the input level. It consists of five feedback loops with time-constants in the range from 5 to 500 ms. For a stationary input signals, the output approaches a logarithmic compression. For rapid input variations the transformation through the adaptation loops is more linear, leading to an enhancement in fast temporal variations, such as onsets. The

output of the adaptation stage is processed by a first-order lowpass filter with a cut-off frequency at 150 Hz, followed by the modulation filterbank, which is a bank of bandpass filters tuned to different modulation frequencies.

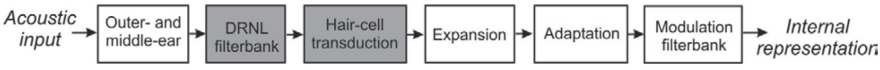


Fig. 2: Structure of the model's front-end. The acoustic input is processed by several stages of auditory processing to form an internal representation with axes; time, frequency and modulation frequency.

Jepsen and Dau (2010) described a method to adjust the parameters of the cochlear stages of the modelto simulate degraded processing due to hair-cell loss. Here, the input/output (I/O) behavior of the DRNL filterbank was adjusted to correspond to the Basilar membrane (BM) I/O functions estimated behaviorally in the HI listeners. The DRNL I/O function was assumed to be linear for levels lower than the lowest measurable point in the data. After parameters were determined at 1 and 4 kHz, linear interpolation and extrapolation were used to obtain parameter-sets for a range of filter center frequencies (0.1 to 8 kHz). The suggested procedure also provided estimates of the effects of OHC and IHC losses with respect to sensitivity. OHC loss was derived from the fitted I/O functions and the loss of sensitivity due to IHC loss was simulated as a linear attenuation at the output of the hair cell transduction stage.

The output of the pre-processing stages, the IR of the respective input stimulus, has the dimensions, time (t), frequency (f) and modulation frequency (mf). IRs were generated using DRNL filters in the range from 0.1 to 8 kHz, with 4 filters per equivalent rectangular bandwidth (60 channels in total) and considering the first six modulation filters, with modulation filter center frequencies ranging from 0 to 46 Hz. In the following, the IRs are defined as the model's response from time 150 to 600 ms.

The back end

Messing *et al*. (2009) introduced the concept of using synthesized DRT diphones and a detector (back end) based on the L_2-norm. With their method, it can be assumed that the source of model errors must be originating from the front-end processing: Here, the same synthesized diphones and the same detector were used. Templates (Y) were the IRs of each diphone presented in a random realization of the noise, at a SNR of 5 dB. For a given test diphone, the IR was calculated at a particular SNR (IR_x), and the mean-squared-errors (MSEs) between IR_x and the two possible templates (e.g., for /*daunt*/ and /*taunt*/) was calculated across time, frequency and modulation frequency. The MSE represents the L_2-norm or Euclidean distance and is considered here as representing the *perceptual* distance between test IR and template. The MSE between IR_x and the correct template (Y_C) was denoted MSE_C, while the MSE calculated for the corresponding wrong template (Y_W) was denoted MSE_W:

$$MSE_C(x) = \frac{\sum_{t=1}^{N_T}\sum_{f=1}^{N_F}\sum_{mf=1}^{N_{MF}}[Y_C(t,f,mf) - IR_x(t,f,mf)]^2}{N_T N_F N_{MF}} \quad \text{(Eq. 1)}$$

$$MSE_W(x) = \frac{\sum_{t=1}^{N_T}\sum_{f=1}^{N_F}\sum_{mf=1}^{N_{MF}}[Y_W(t,f,mf) - IR_x(t,f,mf)]^2}{N_T N_F N_{MF}} \quad \text{(Eq. 2)}$$

where t represents the time index, f is frequency index and mf is the modulation frequency index. N_T, N_F and N_{MF} represent the total number of time, frequency and modulation frequency indices, respectively. The argument x indicates that this calculation is carried out for each of the 192 diphones. Detection is based on the difference, $\Delta MSE = MSE_W - MSE_C$. For $\Delta MSE < 0$, a wrong decision was made, while for $\Delta MSE > 0$, the correct word was detected. In the present study, a probabilistic decision criterion was introduced which reflects the introductions of internal noise in the model. In this approach, the probability of being correct followed a Gaussian cumulative density function with $\mu = 0$ and standard deviation σ.

RESULTS

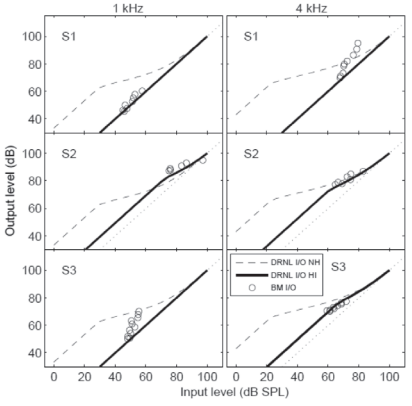
Measured input-output functions

BM I/O functions were derived from the TMC data. The circles in the left side of Fig. 3 represent estimated I/O functions for the three listeners. These were obtained by using the method similar to Jepsen and Dau (2010): A straight line was in each case fitted to the off-frequency masking data, which reflects the masker level at signal threshold as a function of the masker-signal separation. These output levels were then plotted as a function of the input level corresponding to the masker-signal intervals measured with on-frequency masking. The dotted line has a slope of one indicating a linear I/O behavior.

The slopes of the estimated I/O function were calculated and are shown in the right side of Fig. 3. BM compression was found for listeners S2 at both 1 and 4 kHz, and for S3 at 4 kHz, but in all cases slopes were higher than normal (about 0.25). The three other conditions showed a slight expansion (S1 at 1 kHz) or a substantial expansion (S1 at 4 kHz and S3 at 1 kHz). As in Jepsen and Dau (2010), the general observation was that I/O functions were different across listeners even though their sensitivity at the tested frequencies was comparable.

The dashed curves in Fig. 3 (left) shows the DRNL model's I/O function for normal-hearing listeners at the corresponding signal frequencies. Following the procedure suggested in Jepsen and Dau (2010) a set of frequency-dependent DRNL parameters was determined, such that the DRNL I/O function fitted the BM I/O data. In the three cases where the estimated compression exponent was higher than one (expansive), linear I/O functions were assumed. The solid black curves represent the fitted DRNL I/O functions. The sensitivity loss due to OHC loss (HL_{OHC}) and IHC loss (HL_{IHC})

was estimated on the basis of these derived DRNL I/O functions.. The models fitted to listeners S1, S2 and S3 were named M1, M2 and M3, correspondingly. Pure-tone thresholds were predicted and are indicated by the black symbols in Fig. 1. It can be seen that the frequency dependent sensitivity is appropriately accounted for, typically within 10 dB.



Listener	S1	S2	S3
BM compression, 1 kHz	1.34	0.36	2.88
BM compression, 4 kHz	2.01	0.66	0.64
HL_{OHC} at 1 kHz (dB)	35	25	35
HL_{OHC} at 4 kHz (dB)	45	35	35
HL_{OHC} range in filterbank	17-45	10-35	10-45
HL_{IHC} at 1 kHz (dB)	10	20	0
HL_{IHC} at 4 kHz (dB)	10	10	10
HL_{IHC} range in filterbank	0-15	10-35	0-10

Fig. 3: Left: Measured and simulated BM I/O functions for the three HI listeners. Left and right columns show the results at 1 and 4 kHz, respectively. The circles indicate the measured I/O functions for listeners. The dashed curve shows the I/O function for the corresponding DRNL filter for modelling "normal hearing" (MNH). The black curve indicates the DRNL I/O function adjusted to fit the measured I/O function. The dotted line indicates linear I/O behavior. **Right**: Estimated BM compression exponents in units of dB/dB and estimated loss of sensitivity due to OHC and IHC loss in dB.

Figure 4 illustrates how the simulated individual BM compression, frequency selectivity and IHC loss, affected the processing of the speech signals. The shown IRs of the "normal hearing" model (MNH) and three HI models were generated from /*daunt*/ at a SNR of 10 dB. For illustrative purposes, the modulation filterbank was disregarded in this example and replaced by a modulation-lowpass filter with a cut-off frequency at 8 Hz. This representation can be regarded as an auditory spectrogram with axes time and frequency. Darker colors reflect a larger internal excitation. In the IR of M1 it can be observed that the corresponding signal was less resolved in frequency, and the overall amplitude of excitation is reduced (brighter colors). This was expected due to the linear BM processing, and consequently lower sensitivity. For M2 it is clear that most low-frequency (< 500 Hz) information was lost, due reduced sensitivity in these channels. The excitation was also lower for the higher frequencies, and the frequency resolution for the mid- and high frequencies was better than for M1. This reflected the simulated residual compression M2. M3 had a stronger excitation compared to M1 and M2. This was expected since listeners S3 had the mildest hearing loss in terms of sensitivity.

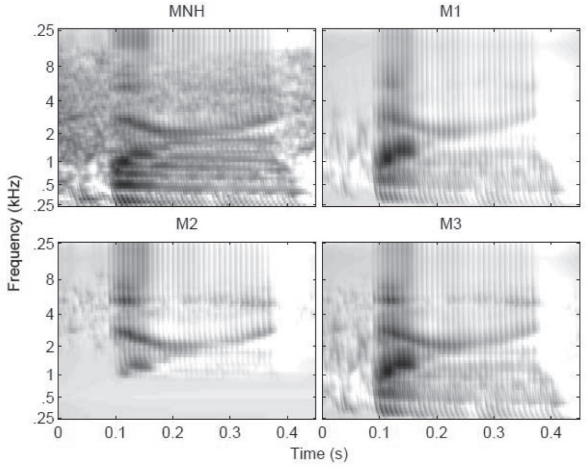


Fig. 4: Examples of IRs of the four models (MNH, M1, M2 and M3).

DRT error patterns

The human performance in the DRT was presented as the error rate percentage in each of the tested acoustic-phonetic dimensions. They are further resolved in two groups, indicating whether the attribute was present (+) or not (-). The chance level in this task was 50% and is indicated by the dashed horizontal line in Fig. 5 (left). The error rates are represented by the bars and shown for NH listeners and the three HI listeners, respectively. These plots are referred to as DRT error patterns. The data for NH listeners were taken from Messing *et al*. (2009). For NH listeners it can be observed that the dimensions VC-, ST+ and ST- have the highest error rates. The remaining dimensions typically have error rates below 20% and for NS the error rate is close to zero errors. The general trend was that more errors were produced at the lower SNR (0 dB). According to Messing *et al*. (2009) these error rates are similar to those measured in NH listeners with natural DRT stimuli (except for NS). All other data were measured in the three HI listeners in the present study. The error bars indicate ± one SD across eight diphone repetitions within the listener. Listener S1 of the present study clearly produce more errors than NH listeners at both SNRs. ST+ has the highest amount of errors of about 60%, exceeding the chance level. In dimensions VC, SB, GV and CM the amount of errors are approximately doubled or more. It appears that this listener is had very good performance in the NS dimension. Overall, there are slightly fewer errors at the higher SNR (10 dB). S1 had the worst overall performance among the HI listeners of this study. S2 shows a high error rate (above 30%) in all dimensions at the SNR of 0 dB, except for NS. At SNR=10 dB there are substantially fewer errors, so it seems that S2 had a great benefit from the better SNR (except for dimension ST). Interestingly S2 was the only HI listener producing errors in dimension NS higher than 6%. S3 also shows a benefit from the better SNR, and had the best performance at SN =0 dB among the HI listeners.

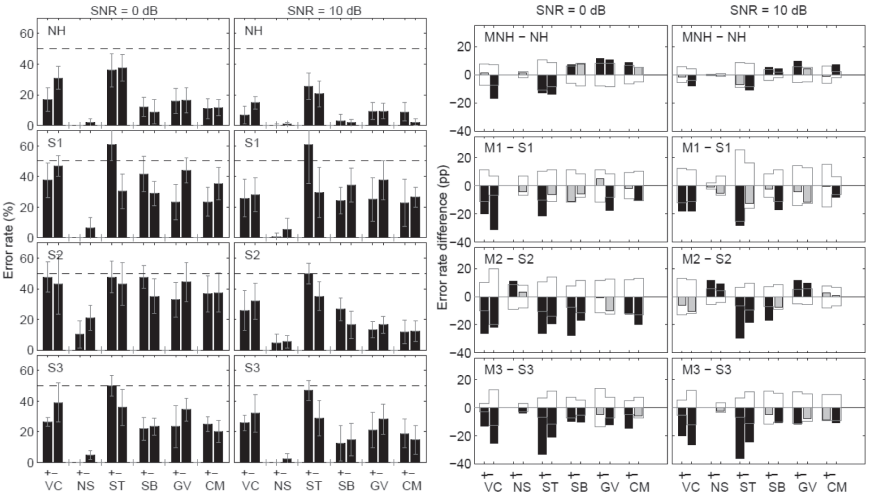


Fig. 5: Left: Measured error patterns of the DRT at the two SNRs (0 and 10 dB). The bars indicate the error rate in percent in the six acoustic-phonetic dimensions, and +/- indicate if the attribute was present or not. Error bars indicate ± one SD. The horizontal dashed line indicates the chance level at 50%. **Right**: Comparison of predicted and measured errors, calculated as (model - human) at SNRs of 0 and 10 dB. Bars represent the error rate difference in percentage points (pp). The boxes indicate ± one SD of the measured errors. Gray bars indicate cases where the error rate is within one SD, while black bars indicate cases where one SD is exceeded.

The target error rates of the models were the human error rates. These are compared in Fig. 5 (right), showing the error rate difference (model - human) in percentage points (pp). Zero error rate differences represent perfect matches. Positive differences reflect that the model overestimates the amount of errors. The boxes indicate ± one SD of the human data. It was considered a good match to human if the predicted error rate was within one SD, and gray bars indicate those cases, while the black bars indicate cases where one SD was exceeded. A χ^2-statistic was used to evaluate if the match between model and human error rate was significant. It evaluates whether error rates were statistically similar or different. Lower values of the χ^2-statistic reflect closer matches. If the χ^2-value is lower than a critical value it cannot be rejected that the errors rates are the same. For one degree of freedom these critical value at a confidence level of 99% was 6.64. In the present study χ^2-values were calculated as the mean across the acoustic-phonetic conditions, excluding dimensions VC-, ST+ and ST-, since these particular conditions were substantially underestimated by all models.

The match between MNH and NH had χ^2-values were 4.40 and 4.26 for SNRs at 0 and 10 dB, respectively. In the case of M1-S1 the χ^2-value was 3.60 for the 0 dB SNR and 4.27 for the SNR of 10 dB, both within the critical limit. For M2-S2 the performance the χ^2-values were 8.17 and 4.54, at 0 and 10 dB SNRs, respectively. Thus for M2 the

performance at the 0 dB SNR condition exceeded the critical value and could not be considered to be a good match. For M3-S3 the SNR of 0 dB error rates at VC+ and CM+ where highly underestimated and $\chi^2 = 6.47$, thus marginally within the critical limit. For the SNR = 10 dB VC+ error were underestimated. The χ^2-value here was 5.61, thus within the critical limit. Overall, all matches are, on average, within the limit in the χ^2 statistical metric, except for M2-S2 at SNR = 0 dB.

DISCUSSION

The predicted error patterns were in reasonable agreement with the measured error patterns, except for a few particular acoustic-phonetic dimensions. This indicates that the front-end processing, which was fitted to the hearing loss of the individual listeners, appropriately simulates internal representations in the hearing-impaired listeners. The IR reflects aspects of both sensitivity and supra-threshold deficits. These effects are crucial for describing the variability in results of HI listeners. The underlying assumption in the detector was that the dominant source of errors produced by the model was due to insufficient information to discriminate a diphone pair, after front-end processing. The use of the synthesized DRT stimuli, combined with the template matching paradigm, allowed this assumption.

The error rates of the VC-, ST+ and ST- dimensions were substantially underestimated. This was also a general problem in the model predictions of Messing *et al*. (2009). These are also the conditions where the NH listeners had substantially worse performance in the DRT using synthetic, compared to natural diphones (Messing *et al*., 2009). The missing capability to predict these error rates may reflect that the detector of the present model is too sensitive to timing cues. Voicing and sustention are the features which primarily depend on timing-differences in a minimal pair, while the remaining features depend more on spectral or spectro-temporal differences.

REFERENCES

Ghitza, O. (**1993**). "Adequacy of auditory models to predict human internal representation of speech sounds," J. Acoust. Soc. Am. **93**, 2160-2171.

Jepsen, M., Ewert, S. D., and Dau, T. (**2008**). "A computational model of human auditory signal processing and perception," J. Acoust. Soc. Am. **124**, 422-438.

Jepsen, M. L., and Dau, T. (**2010**). "Characterizing auditory processing and perception in individual listeners with sensorineural hearing loss," *submitted* to J. Acoust. Soc. Am.

Messing, D. P., Delhorne, L., Bruckert, E., Braida, L. D., and Ghitza, O. (**2009**). "A non-linear efferent-inspired model of the auditory system; matching human confusions in stationary noise," Speech Commun. **51**, 668-683.

Voiers, W. D. (**1983**). "Evaluating processed speech using the diagnostic rhyme test," Speech Technol. **1**, 30-39.