Predicting effects of hearing-instrument signal processing on consonant recognition and confusions

Johannes Zaar[1] and Torsten Dau[1]

[1] *Hearing Systems group, Technical University of Denmark, DK-2800 Kgs. Lyngby, Denmark, Email: jzaar@elektro.dtu.dk*

Introduction

To better understand how various aspects of hearing-instrument processing affect the fundamental speech cues, computational models of speech perception may provide useful information about the auditory cues that contribute to the recognition of a specific consonant or its confusion with another consonant. Recently, the authors of the current study proposed a consonant perception model [1], which combines an auditory model [2] with a temporally dynamic correlation-based template-matching back end. The model was evaluated using the extensive data set from [3], obtained in NH listeners with consonant-vowels (CV) syllables presented in white noise at various signal-to-noise ratios (SNRs) and shown to account well for consonant recognition and consonant confusions.

Several studies have investigated the effects of hearing impairment and hearing-aid (HA) amplification on consonant perception, e.g. [4, 5]. Schmitt *et al.* (2016) [6] presented a consonant perception test specifically designed for high-frequency HA fitting and demonstrated that the test was sensitive to effects of high-frequency amplification as well as to effects of nonlinear frequency compression (NLFC), which is designed to restore high-frequency acoustic information in listeners with pronounced high-frequency hearing loss by compressing the high-frequency signal content and shifting it to lower frequencies. However, NLFC with "too strong" settings can result in a drastic reduction of consonant recognition [6], as frequency-compressed high-frequency consonants may perceptually "morph" into other consonants. In addition to such spectral modifications induced by NLFC, temporal signal modifications induced by highly nonlinear processing schemes typically applied in HAs (e.g., impulse-noise suppression, INS) may also affect consonant perception.

An alternative compensation strategy is represented by cochlear-implant (CI) processing, applied in more severe cases of hearing impairment, using an implanted electrode array. However, CIs are limited with respect to spectral resolution (for review see [7]). DiNino *et al.* (2016) [8] investigated the effect of CI processing with poor electrode-neuron interfaces on the perception of consonants and vowels in NH listeners using vowel-consonant-vowel (VCV) and consonant-vowel-consonant (CVC) syllables, respectively, noise-vocoded to simulate CI processing. Energy from different frequency regions was either zeroed out or redistributed to neighbouring channels, inducing considerable perceptual differences across conditions in the vowel perception test, whereas the consonant perception test showed less variability across conditions.

The present study investigated the predictive power of the model by Zaar and Dau (2017) [1] in several HA and CI processing conditions.

Method

Experiment 1: Effects of HA signal processing

The speech material was taken from the speech material recorded by Schmitt *et al.* (2016) [6] and consisted of the VCVs /aba, aga, ada, apa, aka, ata, asa, aʃa, afa, atsa/[1], spoken by a female native German speaker. Two differently spectrally shaped versions of /asa/ (/asa6/ and /asa9/) and /aʃa/ (/aʃa3/ and /aʃa5/) were defined in [6]. The initial vowels of the considered VCV tokens were manually removed to obtain the CVs /ba, ga, da, pa, ka, ta, sa6, sa9, ʃa3, ʃa5, fa, tsa/.

Five conditions were considered: *unaided*, *default*, *NLFC*, *INS*, and *NLFC&INS*. The unaided condition was a natural listening situation. For the other four conditions, Phonak Naida V90-RIC HAs were employed, assuming a moderate to severe hearing loss. The *default* condition was defined as the default HA settings suggested by the fitting software. In the *NLFC* condition, the strongest possible setting of the provided NLFC algorithm (Phonak SoundRecover) was selected. In the *INS* condition, the strongest possible setting of the provided INS algorithm (Phonak SoundRelax) was selected. In the *NLFC&INS* condition, NLFC and INS were combined using the respective strongest possible settings.

One sound file with all CVs was obtained by concatenating the CVs with 500-ms pauses between them. Steady-state speech-shaped noise (SSN) was added at an effective SNR of 8 dB. 10 seconds of noise alone preceded the first CV. The mixture of CVs and noise was played back frontally from a loudspeaker to a KEMAR dummy head in a sound-attenuating room (speech level: 70 dBA) and the signals were recorded at the position of the dummy head's tympanic membrane. The recordings were equalized to compensate for the applied amplification and cut into the individual CV stimuli.

Ten adult NH native German listeners (mean age: 29.5 years) were tested. The listeners were seated in a sound-insulated booth and binaurally presented with the diotic stimuli via Sennheiser HD 650 headphones at 60 dB sound pressure level. They were asked to select the consonants they heard on a graphical user interface. Each of

[1] Only the subset /ada, aha, ama, aka, asa, aʃa, afa/ of the recorded VCVs were eventually used in [6]. The present study used a different subset.

the 60 stimuli (12 CVs in five conditions) was presented 8 times to each listener, in randomized order. The data obtained for each stimulus were pooled across listeners (80 observations per stimulus).

Experiment 2: Effects of CI signal processing

DiNino *et al.* (2016) [8] considered sixteen VCVs, consisting of consonants embedded in an /aCa/ context (/p/, "apa"; /t/, "ata"; /k/, "aka"; /b/, "aba"; /d/, "ada"; /g/, "aga"; /f/, "afa"; /θ/, "atha"; /s/, "asa"; /ʃ/, "asha"; /v/, "ava"; /z/, "aza"; /dʒ/, "aja"; /m/, "ama"; /n/, "ana"; /l/, "ala"). All VCVs were spoken by a male talker (native speaker of American English). Noise-vocoder processing was applied to the stimuli to simulate CI processing in combination with regions of poor neural survival, using CI simulation software developed by Litvak *et al.* (2007) [9] (15 vocoder bands with logarithmic spacing between 250 Hz and 8.7 kHz). As a control condition, the VCVs were processed using all vocoder bands (*AllChannels*). For the other six conditions, the spectral information in three frequency regions (*Apical* / 421 – 876 Hz; *Middle* / 877 – 1826 Hz; *Basal* / 1827 – 3808 Hz) was degraded by either (i) setting the corresponding channels to zero (*Zero*) or (ii) setting them to zero and adding half of the envelope energy from the zeroed channels to the neighboring lower-frequency channels and the other half to adjacent higher-frequency channels (*Split*). Twelve adult NH listeners with a mean age of 25.2 years participated in the study (native speakers of American English). All 112 VCV stimuli (16 VCVs × 7 conditions) were frontally presented 6 times to each listener at 60 dBA via a loudspeaker in a sound-insulated booth. The data obtained for each stimulus were pooled across listeners (72 observations per stimulus).

Model simulations

The consonant perception model of Zaar and Dau (2017) [1] was used to predict the perceptual data obtained with the HA-processed CVs and with the CI-processed VCVs. Figure 1 shows the model, which combines the auditory model front end of Dau *et al.* (1997) [2] (consisting of a gammatone filterbank, an envelope extraction stage, a chain of adaptation loops and a bank of 4 modulation filters) with a temporally dynamic correlation-based back end, cf. [1]. For a given noisy speech signal, the temporal pattern of the noise alone is subtracted from the corresponding temporal pattern of the noisy speech. The resulting model representations of the test signal and of a set of templates are then aligned in time using a dynamic time warping (DTW) algorithm. Finally, the cross-correlation coefficients between the time-aligned test-signal representation and the time-aligned template representations are calculated and, after adding a constant-variance internal noise to limit the model's resolution, converted to response percentages.

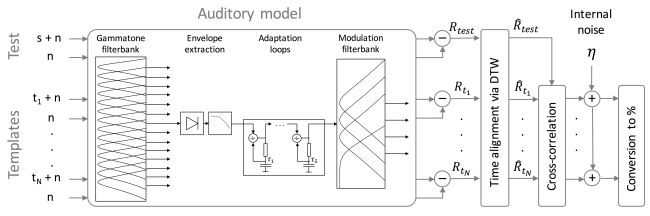

Figure 1: Scheme of the consonant perception model (reprint from [1]).

To predict the data from experiment 1, the experimental stimuli were fed to the model along with "noise alone" signals obtained in the same HA processing condition. The "unaided" stimuli were employed as templates, considering 9 iterations with randomly selected "noise alone" signals for the templates. After obtaining the correlation coefficients between each test signal and all templates, the internal noise was added and the model response for each iteration was defined as the template showing the largest correlation with the test signal. As proposed in [1], the model was calibrated by adjusting the variance of the internal noise based on the average consonant recognition scores obtained all considered conditions. Here, a variance of $\sigma^2_{int,1} = 0.15$ was found to be optimal.

The data from experiment 2, collected by DiNino *et al.* (2016) [8], were predicted in a similar fashion, using the vocoded VCVs in the considered vocoder conditions as test signals and the unprocessed VCVs as templates. In contrast to experiment 1, the experimental stimuli contained no additive noise and the "noise alone" pattern was therefore omitted. Nine iterations of the model simulation were run using newly generated noise-vocoded stimuli in each iteration. An internal-noise variance of $\sigma^2_{int,2} = 0.071$ was found to be optimal based on the average recognition scores obtained in the considered conditions.

Results and discussion

The grand average consonant recognition scores obtained in the five experimental conditions considered in experiment 1 indicated that the consonant recognition was at ceiling (above 90%) for all conditions except the ones including NLFC, namely *NLFC* (55%) and *NLFC&INS* (56%). Only these two conditions were further investigated. To inspect the data more closely in terms of the consonant recognition and confusion scores, Fig. 2 and Fig. 3 show the measured and predicted confusion matrices (CMs) obtained in the *NLFC* and *NLFC&INS* conditions, respectively.

In the *NLFC* condition (Fig. 2) the model provided quite accurate predictions of the stimulus-specific recognition scores, as indicated by the good agreement of the red and gray circles on the "diagonal" of the CM (which has two "steps" as two representations of /s, ʃ/ were considered as stimuli). Furthermore, the model predicted some of the confusions remarkably well (particularly for /d, s6, s9, ts/), although the extent of the confusions was partly underestimated. However, some distinct confusions were not accounted for by the model (/t/ confused with /k/) or predicted to a lesser extent such that they are not visible in Fig 2. For example, /ʃ3, ʃ5/ were confused with /f/, but the predicted response probabilities for /f/ were just below the 7%-threshold. Moreover, the model pre-

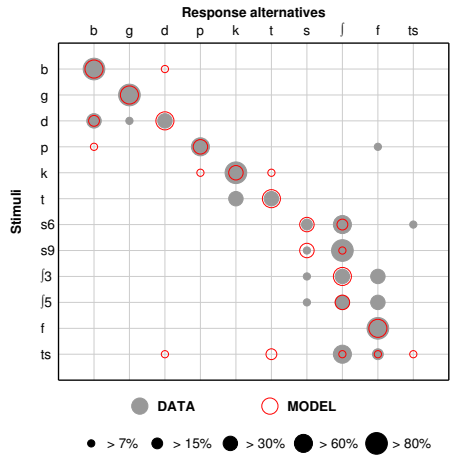

Figure 2: Confusion matrix showing the data and model predictions obtained in the *NLFC* condition of exp. 1.

dicted some additional confusions that were not observed in the perceptual data.

The perceptual data obtained in the *NLFC&INS* condition (Fig. 3) were largely comparable to the data obtained in the *NLFC* condition (Fig. 2). However, some clear differences can be observed (gray circles), as in the *NLFC&INS* condition /k/ was confused with /p, f/ and

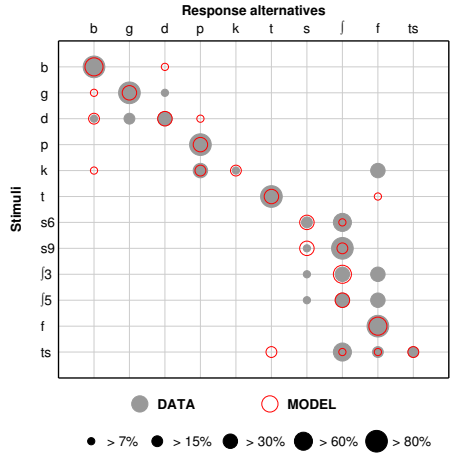

Figure 3: Confusion matrix showing the data and model predictions obtained in the *NLFC&INS* condition of exp. 1.

the confusion of /t/ with /k/ observed in the *NLFC* condition disappeared. Furthermore, /ts/ was not recognized at all in the *NLFC* condition, but was recognized to some extent in the *NLFC&INS* condition. The model predictions captured these perceptual changes between the *NLFC* and the *NLFC&INS* condition well, apart from the confusion of /k/ with /f/, which was not accounted for by the model.

To evaluate the significance of the agreement between the measured and the predicted consonant recognition scores (on-diagonal elements of the CMs), a correlation analysis was conducted, which revealed that the measured and predicted recognition scores were significantly ($p < 0.05$) correlated across stimuli for both the *NLFC* ($r = 0.56$) and the *NLFC&INS* ($r = 0.67$) condition. To furthermore quantify the agreement between the measured and predicted confusions, a correlation analysis of the consonant confusions was performed. For each stimulus, the correlation between the erroneous part of the measured and predicted response patterns (off-diagonal elements of the CMs) was obtained across response alternatives. This analysis was only performed for the stimuli that showed an error of $P_e > 20\%$ in the perceptual data. Table 1 shows the results of the confusion correlation analysis, which revealed that the confusions were positively correlated for all considered stimuli, with most correlations being significant. Note that the large confusion correlations found in the two conditions for /ʃ3, ʃ5/, which are not reflected in Fig. 2 and Fig. 3, are due to model confusion predictions that were qualitatively similar to the measured data but scaled down such that they did not exceed the 7% threshold used for plotting.

Table 1: Pearson's correlation coefficients across response alternatives between measured and predicted consonant confusion patterns obtained in the *NLFC* and *NLFC&INS* conditions of exp. 1. Significant correlations ($p < 0.05$) are given in bold font. The confusion correlation was only obtained for stimuli with a measured error $P_e > 20\%$.

Consonant	*NLFC*	*NLFC&INS*
/b/	–	–
/g/	–	–
/d/	**0.97**	0.25
/p/	0.16	–
/k/	–	**0.62**
/t/	0.12	–
/s6/	**0.94**	**0.93**
/s9/	**0.97**	**0.97**
/ʃ3/	**0.89**	**0.65**
/ʃ5/	**0.88**	**0.78**
/f/	–	–
/ts/	0.25	0.05

As reported by DiNino *et al.* (2016) [8], the grand average consonant recognition scores measured in the seven experimental conditions of experiment 2 were below ceiling and showed little variability across conditions (73% ± 5%) and a large variability across stimuli (with standard deviations of about 30%). The predicted recognition scores exhibited a similar behaviour, albeit with a somewhat smaller variability across stimuli (with standard deviations of about 18.5%).

Figure 4 shows the measured (filled gray circles) and predicted (open red circles) CMs obtained in the *AllChannels* control condition. The main measured confusions were /g/ with /d/, /p/ with /t/, /k/ with /t/, and /th/ with /v/, which resulted in low recognition scores for these stimuli. The main confusions were well accounted for but slightly underestimated by the model, except for /th/ confused with /v/, where the model predicted a perfect recognition of /th/. Thus, the predicted stimulus-specific recognition scores (along the CM's diagonal) showed a similar trend as their measured counterparts, except for the recognition score for /th/. However, the model also predicted some confusions that were not represented in the data. To evaluate the significance of the agreement between the measured and the predicted

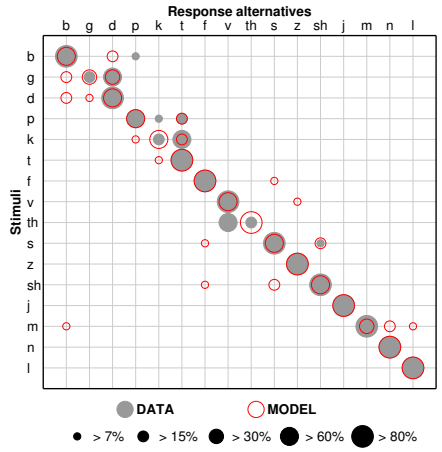

Figure 4: Confusion matrix showing the data and model predictions obtained in the *AllChannels* condition of exp. 2.

consonant recognition scores, a correlation analysis was conducted, which revealed that the measured and predicted recognition scores (on-diagonal elements of the CMs) were significantly ($p < 0.05$) correlated across stimuli for all but the *AllChannels* and *BasalZero* conditions. A correlation analysis of the consonant confusions was performed to also quantify the relation between the measured and the predicted confusions using only the erroneous part of the response patterns (off-diagonal elements of the CMs). As before, this analysis was conducted only for the stimuli that showed a perceptual error of $P_e > 20\%$. Table 2 summarizes the results, which revealed that the confusion correlations for the considered stimuli were very large (mostly above $r = 0.8$) and significant ($p < 0.05$) for the majority of the considered stimuli. However, as observed in the Fig. 4, the /th/ confusions were not well predicted by the model (presumably because they originated from a phoneme-frequency effect rather than from the signal characteristics) and the measured and predicted confusions obtained for /b, d/ in the two *Apical* conditions and for /j/ in the *BasalZero* condition showed either weak correlations or none at all.

Table 2: Pearson's correlation coefficients across response alternatives between measured and predicted consonant confusion patterns obtained in each condition of exp. 2 (AC: *AllChannels*; AZ: *ApicalZero*; AS: *ApicalSplit*; MZ: *MiddleZero*; MS: *MiddleSplit*; BZ: *BasalZero*; BS: *BasalSplit*). Correlation coefficients indicating significant correlation ($p < 0.05$) are given in bold font. The confusion correlation was only obtained for stimuli with a measured error $P_e > 20\%$.

	AC	AZ	AS	MZ	MS	BZ	BS
/b/	–	0.00	-0.04	–	**0.96**	–	–
/g/	**0.92**	**0.87**	**0.90**	**0.88**	**0.93**	**0.95**	**0.86**
/d/	–	0.21	0.38	–	–	–	**0.51**
/p/	**0.96**	**0.93**	**0.98**	**0.97**	**0.94**	**0.93**	**0.87**
/k/	**0.90**	**0.71**	**0.82**	**0.79**	**0.82**	**0.86**	**0.84**
/t/	–	–	–	–	–	–	–
/f/	–	–	–	–	–	–	–
/v/	–	–	–	–	–	–	–
/th/	0.06	-0.11	-0.03	0.38	-0.05	0.08	0.02
/s/	–	–	–	–	–	–	–
/z/	–	–	–	–	–	–	–
/sh/	–	–	–	–	–	**0.95**	**0.96**
/j/	–	–	–	–	–	0.11	–
/m/	–	–	–	**0.50**	**0.68**	–	–
/n/	–	**0.83**	**0.81**	**0.76**	–	**0.90**	**0.81**
/l/	–	–	–	–	–	–	–

Conclusion

The present study evaluated the predictive power of the model of Zaar and Dau (2017) [1] regarding effects of HA and CI signal processing on consonant perception. The model was shown to account for most perceptual effects observed in the data, as the predicted consonant recognition and confusion scores were significantly correlated with their measured counterparts for most conditions. The results indicate that the model can account for supra-threshold effects of hearing-instrument signal processing on consonant perception. This suggests a large potential of the model for evaluating and adjusting such processing schemes, in particular when extended to account for individual hearing impairment.

Acknowledgments

The authors thank Nicola Schmitt and Ralph-Peter Derleth for their support regarding experiment 1 and Mishaela DiNino and Julie Bierer for providing their data and stimuli (experiment 2). This research was funded with support from the European Commission under Contract No. FP7-PEOPLE-2011-290000.

References

[1] Zaar, J. and Dau, T.: Predicting consonant recognition and confusions in normal-hearing listeners. J. Acoust. Soc. Am. 141 (2017), 1051-1064

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

[3] Zaar, J. and Dau, T.: Sources of variability in consonant perception of normal-hearing listeners. J. Acoust. Soc. Am. 138 (2015), 1253-1267

[4] Trevino, A. and Allen, J. B.: Within-consonant perceptual differences in the hearing impaired ear. J. Acoust. Soc. Am. 134 (2013), 607-617

[5] Scheidiger, C., Allen, J. B., and Dau, T.: Assessing the efficacy of hearing-aid amplification using a phoneme test. J. Acoust. Soc. Am. 141 (2017), 1739-1748

[6] Schmitt, N., Winkler, A., Boretzki, M., and Holube, I.: A phoneme perception test method for high-frequency hearing aid fitting. J. Am. Acad. Audiol. 27 (2016), 367-379

[7] Bierer, J. A.: Probing the electrode-neuron interface with focused cochlear implant stimulation. Trends in Amplification 14 (2010), 84-95

[8] DiNino, M., Wright, R. A., Winn, M. B., and Bierer, J. A.: Vowel and consonant confusions from spectrally manipulated stimuli designed to simulate poor cochlear implant electrode-neuron interfaces. J. Acoust. Soc. Am. 140 (2016), 4404-4418

[9] Litvak, L. M., Spahr, A. J., Saoji, A. A., and Fridman, G. Y.: Relationship between perception of spectral ripple and speech recognition in cochlear implant and vocoder listeners. J. Acoust. Soc. Am. 122 (2007), 982-991