Effects of non-stationary noise on consonant identification

JOHANNES ZAAR*, BORYS KOWALEWSKI, AND TORSTEN DAU

Hearing Systems, Department of Electrical Engineering, Technical University of Denmark, Kgs. Lyngby, Denmark

Consonant perception has typically been measured using consonant-vowel (CV) syllables presented in a stationary noise masker at various signal-to-noise ratios (SNRs). Recently, a microscopic speech perception model was proposed (Zaar and Dau, 2017) and shown to account well for consonant perception data obtained in stationary noise. However, unlike stationary noise, real-life interfering sounds typically exhibit strong fluctuations. The present study therefore investigated the effects of highly non-stationary noise on consonant perception and assessed the predictive power of the model in such conditions. Normal-hearing listeners were presented with 15 Danish CVs in 5-Hz interrupted noise at SNRs of −20, −10, 0, and 10 dB. Five different CV onset times with respect to the noise bursts were considered, differing in the amount of induced simultaneous and forward masking. As expected, the consonant recognition scores were inversely related to the amount of simultaneous masking. However, even with minimum simultaneous masking, a substantial loss of consonant recognition was observed at low SNRs, suggesting a forward masking effect. The model, which employs adaptive processes in the front end, accounted for these experimental data to a large extent. The experimental paradigm and the model may be useful for assessing temporal effects of hearing-aid algorithms on consonant perception.

INTRODUCTION

Speech perception has often been measured using sentences as target signals, thus typically providing listeners with context and lexical information that can be exploited to compensate for the sparse acoustic information available in acoustically adverse conditions. To exclude such effects of high-level linguistic processing and, instead, focus solely on the relationship between the available acoustic cues and the speech percept, consonant perception has been measured, typically using consonant-vowel combinations (CVs, e.g., /ta, ba/) at various signal-to-noise ratios (SNRs) in stationary noise (e.g., Miller and Nicely, 1955; Phatak and Allen, 2007; Zaar and Dau, 2015). The resulting consonant recognition and confusion data are useful for investigating the characteristics and confusability of consonant cues. Furthermore, consonant perception tests have been shown to be particularly useful for assessing hearing-aid processing due to the consonants' short-term and high-frequency characteristics (e.g.,

*Corresponding author: jzaar@elektro.dtu.dk

Proceedings of the International Symposium on Auditory and Audiological Research (Proc. ISAAR), Vol. 6: Adaptive Processes in Hearing, August 2017, Nyborg, Denmark. Edited by S. Santurette, T. Dau, J. C.-Dalsgaard, L. Tranebjærg, T. Andersen, and T. Poulsen. The Danavox Jubilee Foundation, 2017.
 ISBN: 978-87-990013-6-1.

Schmitt *et al.*, 2016). Zaar and Dau (2017) proposed a microscopic speech perception model to account for consonant perception data, which combines an auditory processing front end proposed by Dau *et al.* (1997) with a correlation-based, temporally dynamic template-matching back end. The model was shown to account well for the effects of stationary noise (Zaar and Dau, 2017) as well as for spectral effects of hearing-instrument signal processing (Zaar *et al.*, 2017) on consonant recognition and confusions.

In contrast to the stationary masking noise employed in the above-mentioned consonant perception studies, real-life interfering sounds are typically highly non-stationary. While stationary noise introduces only simultaneous masking, non-stationary noise may additionally lead to forward and backward masking of consonant cues. As fine temporal differences presumably play an important role in this context, perceptual effects induced by non-stationary interferers may be particularly useful for evaluating the temporal effects of hearing-aid processing. In the present study, the effect of highly non-stationary noise on consonant identification was measured in normal-hearing (NH) listeners. Special attention was paid to the temporal positioning of the considered CV speech tokens relative to the noise envelope's minima and maxima. Furthermore, the predictive power of the microscopic speech perception model by Zaar and Dau (2017) was evaluated for non-stationary interferers based on the experimental stimuli and the collected data.

EXPERIMENTAL METHOD

Stimuli

The target speech consisted of fifteen consonant-vowel (CV) tokens: /bi, di, fi, gi, hi, ji, ki, li, mi, ni, pi, si, ʃi, ti, vi/ spoken by one male and one female talker (thirty utterances in total). The speech tokens were a subset of the ones employed in a previous study (Zaar and Dau, 2015) and were selected based on maximum intelligibility in stationary noise. The noise was composed of five 100-ms long bursts with 1-ms raised-cosine ramps, separated by 100-ms silent gaps (corresponding to a 5-Hz repetition rate). White noise was chosen as a carrier in order to maximize masking of high-frequency consonants. The presentation level was 65 dB SPL, defined as the level of the noise bursts. Thirty noise waveforms (one per CV utterance) were pre-generated and stored as .wav-files. Each utterance was always presented in combination with the same noise recording. This was done in order to limit the across-repetition variability due to the random fluctuations in the Gaussian noise carrier, whilst preventing noise-learning effects that could occur if only one noise-waveform was used for all utterances (cf. Zaar and Dau 2015). The speech tokens were mixed with the fixed-level noise at four presentation levels: 45, 55, 65, and 75 dB SPL, corresponding to broadband SNRs of -20, -10, 0, and 10 dB. The onsets of the CV tokens were positioned at five different onset times relative to the noise: 400, 450, 500, 525, and 550 ms after the initial noise onset, as shown in Fig. 1. To investigate whether the speech tokens *per se* were sufficiently intelligible at the considered speech levels, two additional conditions with speech in quiet at presentation levels of 45 and 65 dB SPL (termed Q65 and Q45) were considered.

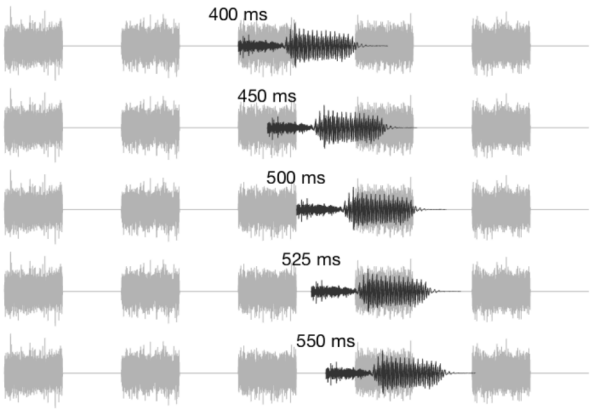

Fig. 1: Stimulus generation. Example of the CV /ki/ (black waveforms) in 5-Hz interrupted noise (gray waveforms) for the five considered CV onset times, as indicated above the respective waveforms.

Listeners and procedure

Twelve young NH native Danish listeners aged 19-26 (average age: 21.7 years) were tested. The normal-hearing status was established based on pure-tone thresholds lower than 20 dB HL in the 250 – 8000 Hz range. The listeners were seated in a sound-attenuating listening booth, monaurally presented with the stimuli via headphones, and asked to indicate the consonant they heard on a graphical user interface. The stimulus presentation could not be repeated and no feedback was provided to the listeners. Six different experimental blocks were defined based on the two quiet conditions and four SNRs (order: Q65, Q45, SNR = 10, 0, −10, −20 dB). A short training run was provided at the beginning of each block. Within each block, the order of presentation was randomized. In each condition, each stimulus was presented to the listeners five times.

MODELING

Model description

The consonant perception model of Zaar and Dau (2017) was considered to predict the perceptual data obtained in the experiment. Figure 2 shows the model, which combines the auditory model front end of Dau *et al.* (1997) with a temporally dynamic correlation-based back end. The auditory model consists of (i) a bank of 15 fourth-order gammatone filters with center frequencies logarithmically spaced between 315 Hz and 8 kHz, (ii) an envelope extraction stage (realized by half-wave rectification and lowpass filtering at 1 kHz), (iii) a chain of five adaptation loops (designed to mimic adaptive properties of the auditory periphery), and (iv) a bank of four modulation filters, implemented as a 2-Hz lowpass filter in parallel with three second-order bandpass filters with a Q-factor of 1 and center frequencies of 4, 8, and

16 Hz, respectively. For a given noisy speech signal, the temporal pattern of the noise alone (after the preprocessing stages) is subtracted from the corresponding temporal pattern of the noisy speech. The resulting model representations of the test signal (R_{test}) and of a set of templates (R_{t_1}, R_{t_2}, …, R_{t_N}) are then aligned in time using a dynamic time warping (DTW) algorithm. Finally, the cross-correlation coefficients between the time-aligned test-signal representation ($\hat{R}_{test}$) and the time-aligned template representations ($\hat{R}_{t_1}$, $\hat{R}_{t_2}$, …, $\hat{R}_{t_N}$) are calculated and, after adding a constant-variance internal noise to limit the model's resolution, converted to response percentages.

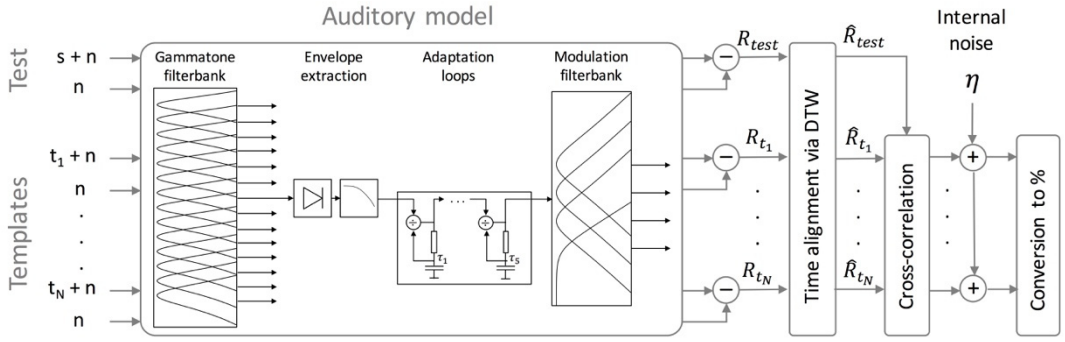

Fig. 2: Scheme of the considered microscopic speech perception model (from Zaar and Dau, 2017).

Simulation procedure

The experimental stimuli employed in the noise conditions were fed to the model as test signals. The templates were created by mixing the fifteen available CV tokens from the test-signal talker with randomly generated interrupted noise, using the same speech level and CV onset time as in the test signal. The "correct" template contained the same speech token as the test signal, whereas the noise signals differed. Randomly generated interrupted noise was considered as "noise alone". This is different from Zaar and Dau (2017), where the noise waveform in the test signals and templates was identical to the "noise alone", and was modified here to simulate potential informational contributions of the noise (i.e., noise bursts being mistaken for consonant cues). Five templates were generated for each speech token, SNR, and CV onset time, each using a different randomly generated interrupted noise waveform. The test signals and templates were passed through the model front end; only the consonant portions of the resulting internal representations, i.e., the portions of the CV tokens between consonant onset and vowel onset, were further considered in the back end. Whereas Zaar and Dau (2017) had considered the entire CV tokens, this modification was necessary here to prevent the model from being influenced by the task-irrelevant vowel portions, in particular when positioned in a noise gap. After obtaining the correlation coefficients between the internal representations of each test

signal and the corresponding templates, the internal noise was added. Consistent with Zaar and Dau (2017), the variance of the internal noise was 0.05 and was held constant across the considered conditions. The model response for each iteration was defined as the template showing the largest correlation with the test signal.

RESULTS AND ANALYSIS

Experimental results

The measured consonant recognition scores were averaged across consonants and talkers. Figure 3 shows the recognition scores in terms of the mean and standard deviations across listeners as a function of speech level for the considered conditions. It can be observed that consonant recognition was at ceiling for the speech tokens presented in quiet (crosses), both for presentation levels of 45 and 65 dB SPL, albeit with a larger standard deviation at 45 dB SPL. Thus, it can be concluded that (i) the speech tokens were perfectly identifiable in quiet and (ii) that audibility was sufficient even at the lowest speech level considered. A two-tailed paired-sample t-test confirmed the latter observation, indicating no significant effect of presentation level in quiet ($p = 0.143$).

The remaining symbols in Fig. 3 depict the recognition scores obtained for the CVs mixed with the fixed-level interrupted noise according to the different CV onset times (cf. Fig. 1). As expected, consonant recognition generally decreased with decreasing speech level. Moreover, a clear effect of CV onset time can be observed. Specifically, the earliest CV onset time of 400 ms resulted in the lowest recognition scores (circles) and increasing CV onset times generally induced increasing recognition scores. However, this trend did not persist for the CV onset time of 550 ms (upward facing triangles), which induced lower recognition scores than the CV onset time of 525 ms (downward facing triangles). Furthermore, the recognition scores obtained for CV onset times of 500 ms and 525 ms were almost identical at the two lowest speech levels. Most of the reduction in recognition scores can be attributed to the degree that the consonant cues were simultaneously masked: As some consonant cues last up to around 100 ms, simultaneous masking was – depending on the CV onset time – induced by the third (CV onset times of 400 and 450 ms; circles and squares) or the fourth (CV onset times of 525 and 550 ms; upward and downward facing triangles) noise burst (cf. Fig. 1). Nonetheless, an effect of forward masking was clearly also present, as the recognition scores obtained in the condition with the least amount of simultaneous masking (CV onset time of 500 ms; diamonds) were much lower than in quiet (crosses) and somewhat lower than in the conditions with more simultaneous masking (CV onset time of 525 and 550 ms; downward and upward facing triangles, respectively). Two-tailed paired-sample t-test comparing all ten combinations of the five CV onset-time conditions were conducted after collapsing the recognition scores across speech level. In accordance with the previous observations, the results indicated highly significant ($p < 0.0001$) differences between all conditions except between CV onset times of 500 and 550 ms ($p = 0.568$). The latter two conditions did, however, exhibit highly significant ($p < 0.0001$) differences at a speech level of 65 dB SPL.

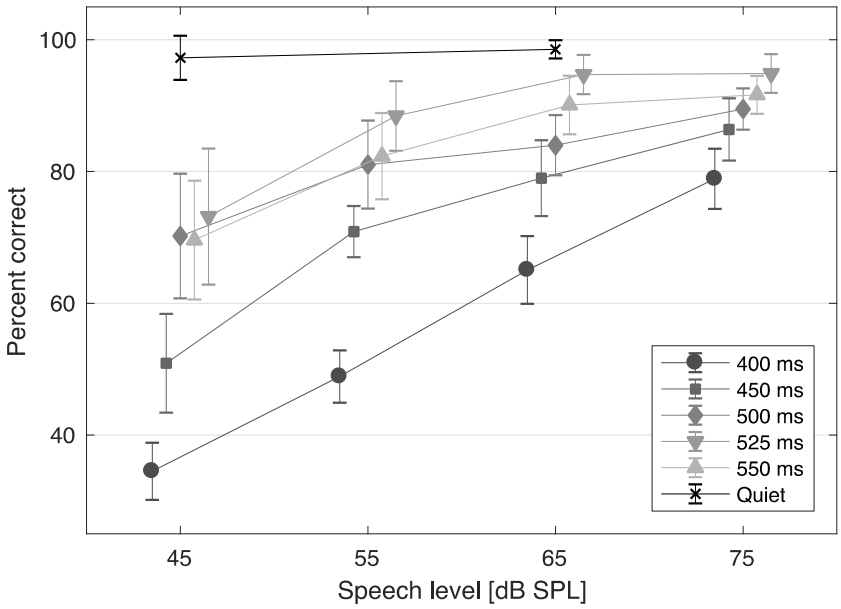

Fig. 3: Average consonant recognition scores as a function of speech level. The crosses represent the quiet conditions. The other symbols represent the different noise conditions, as indicated by the CV onset times in the legend. The noise level was 65 dB SPL. The error bars depict the standard deviation across listeners. A slight horizontal jitter was applied for visual clarity.

Model predictions

The predicted recognition scores obtained for the experimental stimuli in the noise conditions are depicted in the left panel of Fig. 4. Comparing these predictions with the data shown in Fig. 3, it can be observed that the model predictions were generally similar to the measured data as (i) the recognition scores globally decreased with decreasing speech level and (ii) the loss of consonant recognition was proportional to the amount of simultaneous masking. However, the model did not predict the extent of consonant recognition loss induced by the predominantly forward-masking based condition (CV onset time of 500 ms; diamonds), i.e., the effect of forward masking was smaller in the model than measured in listeners. Accordingly, the mean average error between predicted and measured recognition scores was relatively large for the CV onset time of 500 ms (15.8%) and much smaller for the remaining conditions (4.8% on average). Nonetheless, the recognition scores shown in Figs. 3 and 4 (left panel) were overall strongly correlated (Pearson's r of 0.94), as can be seen in the scatter plot presented in the right panel of Fig. 4.

So far, only average consonant recognition scores have been considered. However, different consonants are typically very differently affected by the masking noise (cf. Zaar and Dau, 2015). To investigate whether the model predicted the trends across consonants correctly, the measured and predicted consonant-specific recognition

scores were averaged across speech levels and their Pearson's correlation across consonants was computed. Consistent with the model predictions reported for stationary-noise conditions (Zaar and Dau, 2017), the correlations were large ($r > 0.75$) and highly significant ($p < 0.001$) for all considered conditions except for the CV onset time of 500 ms, for which, nonetheless, significant correlation was found ($r = 0.5$, $p < 0.05$).

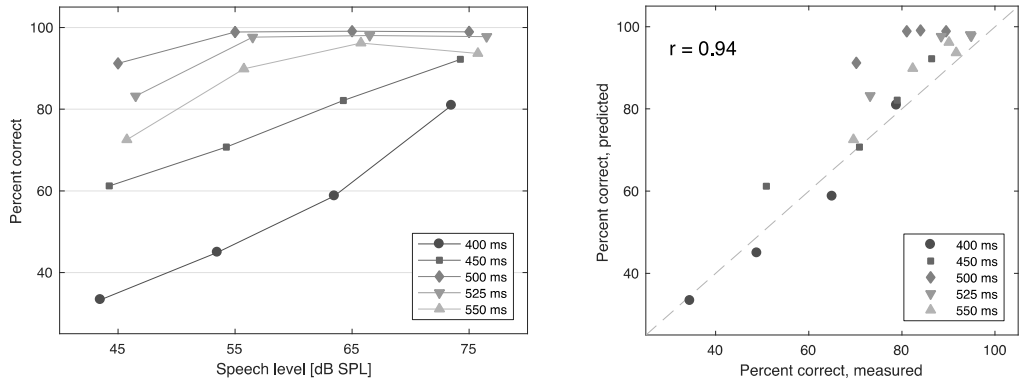

Fig. 4: Left panel: Model predictions of average consonant recognition scores as a function of speech level, corresponding to Fig. 3. Right panel: Model predictions of average consonant recognition scores (shown in the left panel) as a function of their measured counterparts (shown in Fig. 3).

DISCUSSION AND OUTLOOK

The perceptual effects measured in the present study suggest that consonant perception in non-stationary noise strongly depends on the position of the consonant cue relative to the noise envelope's minima and maxima and thus on the amount of simultaneous masking. This is consistent with the well-established observation that listeners make use of "glimpses" of the target speech in fluctuating interferers (e.g., Cooke, 2006). However, since the present study considered CVs as target signals, the experimental paradigm revealed more detailed effects, including a clear effect of forward masking. Thus, the paradigm may be useful for revealing temporal effects of hearing-aid processing, as demonstrated in a related study by Kowalewski *et al.* (2017), which applied the paradigm to investigate the effects of slow- vs. fast-acting compression in hearing-impaired listeners. While only consonant recognition has been discussed here, an additional analysis of the consonant confusions in the data may reveal the interaction between noise and speech tokens in more detail. For instance, it is possible that the noise bursts acted not only as a masker but were even mistaken for consonant cues, thus adding an informational component. It needs to be further investigated, however, whether the detailed effects measured with the considered nonsense speech tokens and artificial interrupted noise also play a role in more realistic conditions.

The model predictions obtained in the present study show that the large predictive power of the model, which had previously been demonstrated in conditions of stationary noise (Zaar and Dau, 2017) and spectral aspects of hearing-instrument processing (Zaar *et al.*, 2017), also extends to conditions of non-stationary noise. Only consonant recognition was considered here and it needs to be investigated whether the model also can account for consonant confusions in the data. Despite the overall accurate predictions, the model was found to be not sensitive enough to the effects of forward masking. While the underlying auditory model (Dau *et al.*, 1997) contains an adaptation stage and does account for "classical" forward-masking data (for narrowband signals), the present speech-based configuration does not seem to fully capture the reported forward-masking effects measured with speech signals. Thus, it may be useful to adapt the model such that it better accounts for this aspect of the data, for instance by modifying the time constants in the adaptation loops or integrating a simulation of the cochlear nonlinearities. Overall, the model may be useful as a tool for analysing temporal effects of hearing-aid processing, in particular when combined with simulations of individual hearing loss.

REFERENCES

Cooke, M. (**2006**). "A glimpsing model of speech perception in noise," J. Acoust. Soc. Am., **119**, 1562-1573. doi: 10.1121/1.2166600

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. doi: 10.1121/1.420344

Kowalewski, B., Zaar J., Fereczkowski, M., MacDonald, E., Strelcyk, O., May, T., and Dau T. (**2017**) "Effects of slow- and fact-acting compression on hearing-impaired listeners' consonant-vowel identification in interrupted noise," Proc. ISAAR, **6**, 375-382.

Miller, G.A., and Nicely, P.E. (**1955**). "An analysis of perceptual confusions among some English consonants," J. Acoust. Soc. Am., **27**, 338-352. doi: 10.1121/1.1907526

Phatak, S.A., and Allen, J.B. (**2007**). "Consonant and vowel confusions in speech-weighted noise," J. Acoust. Soc. Am., **121**, 2312-2326. doi: 10.1121/1.2642397

Schmitt, N., Winkler, A., Boretzki, M., and Holube, I. (**2016**). "A phoneme perception test method for high-frequency hearing aid fitting," J. Am. Acad. Audiol., **27**, 367-379. doi: 10.3766/jaaa.15037

Zaar, J., and Dau, T. (**2015**). "Sources of variability in consonant perception of normal-hearing listeners," J. Acoust. Soc. Am., **138**, 1253-1267. doi: 10.1121/1.4928142

Zaar, J., and Dau, T. (**2017**). "Predicting consonant recognition and confusions in normal-hearing listeners," J. Acoust. Soc. Am., **141**, 1051-1064. doi: 10.1121/1.4976054

Zaar, J., Schmitt, N., Derleth, R.-P., DiNino, M., Arenberg, J.G., and Dau, T. (**2017**). "Predicting effects of hearing-instrument signal processing on consonant perception," J. Acoust. Soc. Am., *under review*.